INTELIGENCIA ARTIFICIAL Y JUSTICIA PENAL: MÁS ALLÁ DE LOS RESULTADOS LESIVOS CAUSADOS POR ROBOTS[1]

Fernando Miró Llinares

Catedrático de Derecho Penal y Criminología de la Universidad Miguel Hernández de Elche

Resumen: El presente trabajo aborda las implicaciones del uso de sistemas de Inteligencia Artificial en la justicia penal. Más allá de los resultados lesivos causados por máquinas, se analiza primero qué es la IA, cuáles son sus características y tipologías y sus usos en relación con la justicia criminal y cuál se prevé que sea su evolución. Se analizan después los principales riesgos éticos que plantea con especial interés por

[1] Este artículo ha sido desarrollado en el marco del proyecto «Criminología, evidencias empíricas y Política criminal. Sobre la incorporación de datos científicos para la toma de decisiones en relación con la criminalización de conductas» — *Referencia: DER2017-86204-R, financiado por la Agencia Estatal de Investigación (AEI)/Ministerio de Ciencia, Innovación y Universidades y la Unión Europea a través del Fondo Europeo de Desarrollo Regional-FEDER— «Una manera de hacer Europa»*. La realización de este trabajo también ha sido posible gracias a la financiación del proyecto PERICLES (*Policy recommendation and improved communication tolos for law enforcement and security agencies preventing violent radicalisation*) con referencia No 740773, del programa Horizonte 2020 de la Unión Europea.

lo que supone para la privacidad y por la denominada «discriminación algorítmica».

Palabras clave: Inteligencia Artificial, *Big Data*, *Machine Learning*, policía predictiva, sistema de justicia penal, discriminación algorítmica, sesgos algorítmicos.

Abstract: This paper addresses the implications of the use of Artificial Intelligence systems in criminal justice. Beyond the harmful results caused by machines, we first analyse what AI is, what their characteristics and typologies are and their uses regarding criminal justice, and what its evolution is expected to be. We then analyse with special interest the main ethical risks it poses concerning privacy and the so-called «algorithmic discrimination».

Keywords: Artificial Intelligence, Big Data, Machine Learning, predictive policing, criminal justice system, algorithmic discrimination, algorithmic bias.

I. No sólo T-800, también, y sobre todo, Skynet

Pese a que vivimos ya inmersos en el ciberespacio, mantenemos relaciones económicas, sociales e íntimas dentro de él, y comenzamos a explorar la realidad virtual como un entorno en el que efectivamente sentimos y actuamos, los seres humanos parecemos seguir instalados en la creencia de que sólo lo físico es real, de que únicamente aquello que es corpóreo y geográficamente ubicable merece ese calificativo o, cuanto menos, de que lo que no puede tocarse con las manos no puede ser nombrado del mismo modo que aquello que sí. Llamamos a Internet espacio virtual, es decir, no real, como si una calumnia hecha pública a través de YouTube fuera menos real que aquella otra enunciada delante de la víctima y de tres personas más; identificamos los juegos *online* como meras simulaciones, obviando el creciente número de conductas de acoso allí realizadas[2], o la utilización de estos entornos por grupos terroristas para la radicalización de jóvenes vulnerables[3]; y cuando nos referimos a los peligros de la Inteligencia Artificial (en adelante, IA) solemos pensar en un robot que obvia su programación para causar daños a personas. Pero cuando lo hacemos no estamos pensando en «Skynet», sino en el robot humanoide «T800», con el aspecto de Arnold Schwarzenegger, o en el «T1000», más maleable pero también corpóreo. Y ello porque Skynet no es más que un algoritmo informático sin forma física, y esto parece cau-

[2] Tang, W. Y., y Fox, J.: «Men's harassment behavior in online video games: Personality traits and game factors», en *Aggressive Behavior*, vol. 42, n.º 6, 2016, pp. 512 y ss.

[3] Al-Rawi, A.: «Video games, terrorism, and ISIS's Jihad 3.0», en *Terrorism and Political Violence*, vol. 30, n.º 4, 2018, pp. 740 y ss.

sarnos menos temor, aunque en la serie de películas de culto que empezó en los años 80 sea precisamente tal IA la que, aprovechándose del control computacional de todas las redes informáticas globales que los propios humanos le dan con la intención de acabar con un falso virus informático que ha infectado e incapacitado Internet, decida acabar con la vida humana y crear otros robots.

Cuando hablamos de IA pensamos en Robótica, pese a que una y otra sean disciplinas distintas, con orígenes distintos, aunque claramente convergentes en el presente y sobre todo en el futuro. Y cuando nos referimos a la relación entre el Derecho penal y la IA solemos preocuparnos por la respuesta del sistema de atribución de responsabilidad penal ante los cursos causales desviados producidos por máquinas con forma física, generalmente con gran potencial destructivo, como robots militares o coches de conducción autónoma, programadas por humanos y con mayor o menor capacidad de autonomía[4]. Pero nos olvidamos de los otros tipos de máquinas, las computacionales, y de los antecedentes de esa exageración cultural que es «Skynet»: los algoritmos informáticos realizados a partir de técnicas matemáticas de procesamiento de información en que consiste la IA[5]. Y lo cierto es que tales sistemas también deben ocuparnos y preocuparnos, pues, pese a no disponer de una forma física concreta, hay muchas formas de IA que inciden en nuestra vida, determinando algunas de nuestras decisiones y muchas de las que nos afectan en aspectos muy variados, afectando a intereses individuales y colectivos dignos de protección.

El presente trabajo aborda la problemática de la IA con la intención de obtener una imagen genérica sobre la relación entre la misma y el sistema de justicia penal. Lo hace, sin embargo, obviando

---

[4] Véase Quintero Olivares, G.: «La robótica ante el Derecho penal: el vacío de respuesta jurídica a las desviaciones incontroladas», en *Revista Electrónica de Estudios Penales y de la Seguridad*, n.º 1, 2017; De la Cuesta Aguado, P. M.: «La ambigüedad no es programable: racionalización normativa y control interno en inteligencia artificial», en *Revista de Derecho y Proceso Penal*, n.º 44, 2016, pp. 165-194.

[5] Hay excepciones. La más significativa es el trabajo de Valls Prieto, especialmente relacionado con el uso de las técnicas de Big Data y su afectación al interés jurídico privacidad. Véanse al respecto tanto, Valls Prieto, J.: *Problemas jurídico penales asociados a las nuevas técnicas de prevención y persecución del crimen mediante inteligencia artificial*, Dykinson, Madrid, 2017; como Valls Prieto, J.: «El uso de inteligencia artificial para prevenir las amenazas cibernéticas», en Valls Prieto, J. (Coord.): *Retos jurídicos por la sociedad digital*, Aranzadi, Navarra, 2018, pp. 77-106. También recientemente, hasta el punto de que han sido incorporadas en fase de revisión de este artículo gracias a la oportunidad dada por los revisores y editores, han afrontado la cuestión Romeo Casabona, C.M: «Riesgo, procedimientos actuariales basados en Inteligencia Artificial y medidas de seguridad», en *Revista Penal,* n.º42, julio 2018, pp. 165 y ss., centrándose especialmente en las herramientas (automatizadas y no automatizadas) de valoración del riesgo, y Nieva Fenoll, J.: *Inteligencia judicial y proceso judicial,* Marcial Pons, Ediciones Jurídicas y Sociales, Madrid, 2018; ocupándose muy particularmente, aunque no sólo, de los aspectos procesales de la inteligencia artificial.

a propósito la cuestión de la responsabilidad penal en relación con daños causados por sistemas de IA, pese a que algunas de las consideraciones que se hacen en este trabajo sobre los tipos de IA y sus características pueden ser útiles en relación con ello y son planteadas como base para futuros análisis. La razón no es el desinterés, pues se trata de una problemática apasionante y necesaria de análisis, sino la convicción de que a la Ciencia Penal, en un sentido amplio, también le debe interesar la utilización de estos sistemas de IA para la prevención, la investigación y la determinación judicial de delitos por la posible afectación a intereses dignos de tutela como la privacidad que puede suponer, así como por los efectos positivos y negativos que la implementación de estos sistemas podría conllevar. Y como toda problemática que está «a caballo» entre diferentes ciencias relacionadas con la justicia penal, pues tiene que ver con los principios y fundamentos de la intervención del Derecho penal, con el sistema constitucional y sus derechos fundamentales en el que se enmarcan aquellos, con la propia legislación procesal que fija las condiciones de su uso o con la administrativa al relacionarse con la acción policial fuera de la investigación del delito, y con la propia ciencia social criminológica que le da sentido a su utilización policial; el análisis y la determinación de las mejores condiciones para su uso corre claramente el riesgo de quedarse en tierra de nadie. La utilización de la IA en el sistema de justicia penal ya es, sin embargo, una realidad que debe ser entendida como tierra de todos y resulta esencial que afrontemos, de un modo casi holístico como se pretende en este trabajo, los retos que plantea su uso y la determinación de los límites y buenas prácticas que deben conformar tal aplicación.

II. La IA y su interés para el Derecho penal

2.1. Tras el término IA: Precisiones conceptuales en torno a una sinonimia parcial

En cierta medida el término IA es una mezcla de eufemismo y desiderátum. Eufemismo porque si bien no hay nada de problemático en la expresión «sistemas para el tratamiento y análisis automático de la información», que es de aquello de lo que realmente estamos hablando, tampoco hay en ello nada evocador como sí lo hay en la expresión «Inteligencia Artificial». Y desiderátum porque detrás de esos sistemas hay algo más que el tratamiento de información, hay una voluntad de que sean inteligentes, capaces de tener o imitar procesos cognitivos propios de los seres humanos, y ello pese a que todavía no conocemos bien el funcionamiento real de la mente ni tampoco está claro qué es eso de la inteligencia humana. Aunque los orígenes se remontan a los años 50 del

pasado siglo[6], actualmente lo que hay bajo la denominación IA es una infinidad de técnicas avanzadas de procesamiento matemático de datos. Entre las principales, caben destacar determinados procesos de *Big Data* para la eficiente gestión de grandes volúmenes de datos[7]; las técnicas de *Data Mining*, que permiten encontrar patrones y resumir grandes volúmenes de datos de forma comprensible y útil, facilitando la toma de decisiones[8]; las de *Machine Learning*, que tienen como finalidad el aprendizaje de las máquinas a medida que incorporan datos actualizados[9]; las del Procesamiento del Lenguaje Natural (PLN) que pretende que los sistemas informáticos entiendan y manipulen el lenguaje de forma natural logrando que la máquina reconozca la voz humana y sea capaz de darle una respuesta lógica[5]; y la visión por ordenador para adivinar, por ejemplo, las emociones que experimenta una persona a partir del reconocimiento facial[6]. Desde entonces, los avances de las nuevas tecnologías han impulsado la evolución de estas técnicas, no tanto conceptualmente sino en la capacidad de procesamiento, dando lugar a nuevas metodologías más avanzadas basadas en la idea de la automatización, como el *Deep Learning*[10] o las *Redes Neuronales Artificiales*[11]. Ambas técnicas están basadas en lo que ya sabemos de la estructura biológica del cere-

---

[6] Benko, A., y Lányi, C. S.: «History of artificial intelligence», en *Encyclopedia of Information Science and Technology, Second Edition*, IGI Global, 2009, pp. 1759-1762; Buchanan B. G.: «A (very) brief history of artificial intelligence», en *AI Magazine*, vol. 26, n.º 4, 2005, pp. 53 y ss.; McCorduck, P.: *Machines who think: A personal inquiry into the history and prospects of artificial intelligence*, AK Peters/CRC Press, 2009.

[7] Labrinidis, A., y Jagadish, H. V.: «Challenges and opportunities with big data», en *Proceedings of the VLDB Endowment*, vol. 5, n.º 12, 2012, pp. 2032-2033; McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., y Barton, D.: «Big data: the management revolution», en *Harvard business review*, vol. 90, n.º 10, 2012, pp. 60-68; Sagiroglu, S., y Sinanc, D.: «Big data: A review», en *Collaboration Technologies and Systems (CTS), 2013 International Conference on*, IEEE, 2013, May, pp. 42-47.

[8] Han, J., Pei, J., y Kamber, M.: *Data mining: concepts and techniques*, Elsevier, USA, 2011; Tan, P. N.: *Introduction to data mining*, Pearson Education, India, 2007.

[9] Michalski, R. S., Carbonell, J. G., y Mitchell, T. M. (Eds.): *Machine learning: An artificial intelligence approach*, Springer-Verlag, Berlín, 2013; Russell, S. J., y Norvig, P.: *Artificial intelligence: a modern approach*, Pearson Education Limited, Malasya, 2016; Marsland, S.: *Machine learning: an algorithmic perspective*, Chapman and Hall/CRC, New York, 2011.

[5] Chowdhury, G. G.: «Natural Language Processing», en *Annual Review of Information Science and Technology*, vol. 37, n.º 1, 2003.

[10] [6] Szeliski, R.: *Computer Vision: Algorithms and Applications.* (Recuperado de: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.414.9846&rep=rep1&type=pdf. Última visita en 05/09/2018)

[7] Martín, B., y Sanz, A.: *Redes neuronales y sistemas borrosos,* Editorial Ra-Ma, Zaragoza, 2006.

Goodfellow, I., Bengio, Y., y Courville, A./Bengio, Y.: *Deep learning*, Vol 1, MIT press, Cambridge, 2016; LeCun, Y., Bengio, Y., y Hinton, G. «Deep learning», en *Nature*, vol. 521, n.º 7553, 2015, pp. 436 y ss.

[11] Schmidhuber, J.: «Deep learning in neural networks: An overview», en *Neural networks*, vol. 61, 2015, pp. 85-117; Haykin, S.: *Neural networks: a comprehensive foundation*, Prentice Hall PTR, NJ, 1994; Hilera González, J. R., y Martínez Hernando, V. J.: *Redes*

bro humano y buscan la adaptación y reproducción de algunas de sus capacidades[7].

La combinación de algunas de estas técnicas en diferentes tipos de máquinas con distintas formas y utilidades han dado lugar a IA muy distintas entre sí. Existen, sin embargo, dos aspectos esenciales que van a determinar el alcance potencial de cada IA, tanto de las que ya existen como de las que se podrían crear virtualmente en el futuro: (1) su capacidad para ejecutar un mayor o menor abanico de instrucciones, y (2) el grado de autonomía con el que las ejecute frente a la influencia del ser humano.

Respecto a la primera, a las posibilidades funcionales de la IA, no me refiero tanto a las utilidades concretas, porque éstas pueden ser tantas como ámbitos en los que se desee aplicar, ya sea para procesar grandes cantidades de datos, predecir patrones, automatizar pequeñas acciones, o estimar probabilidades, sino a la capacidad de razonamiento de la máquina en el sentido del grado de equivalencia entre la complejidad de procesamiento que realiza la máquina y el que es capaz de realizar el cerebro humano. Pues bien, aquí es donde el eufemismo es más evidente, puesto que pese a que hay quienes defienden que en un futuro próximo se darán las posibilidades apropiadas para elaborar sistemas artificiales capaces de asimilar todos los procesos mentales humanos, o incluso generar otros diferentes[12], lo cierto es que hoy las IA funcionan básicamente como modelos matemáticos, y sólo se parecen a la inteligencia humana en la capacidad de cálculo[13]. Pese a que, definitivamente, tengan mayores capacidades que los seres humanos en ciertos campos, estas máquinas no tienen capacidad para el razonamiento complejo y apenas sí para el aprendizaje[14]. Básicamente las máquinas no son capaces, al menos por ahora, de interpretar contextos complejos o de introducir por sí mismas variables nuevas, sino que sólo toman decisiones a partir de premisas lógicas básicas que son previamente introducidas o consideradas. Quién sabe si en un futuro nos podríamos encontrar con sistemas informáticos capaces de aprender autónomamente del entorno y adaptarse a él, acercándose algo a la idea de conciencia[15], pero ahora mismo esta es una realidad utópica.

---

neuronales artificiales: fundamentos, modelos y aplicaciones, RA-MA Editorial, Madrid, 2000.

[12] HOLLAND, O.: «The future of embodied artificial intelligence: Machine consciousness?», en LIDA, F., PFEIFER, R., STEELS, L., y KUNIYOSHI, Y. (EDS.): *Embodied artificial intelligence,* Springer, Berlin, 2004, pp. 37-53.

[13] BENÍTEZ, R., ESCUDERO, G., KANAAN, S., y RODÓ, D. M.: *Inteligencia artificial avanzada*, Editorial UOC, Barcelona, 2014.

[14] LIDA, F., PFEIFER, R., STEELS, L., y KUNIYOSHI, Y. (EDS.): *Embodied artificial intelligence,* Springer, Berlin, 2004, pp. 1-26.

[15] MCDERMOTT, D.: «Artificial intelligence and consciousness», en *The Cambridge Handbook of Consciousness*, 2007, pp. 117-150.

El segundo aspecto esencial para comprender la IA, pero íntimamente relacionado con el anterior, es la cuestión del nivel real de autonomía en la toma de decisiones que puede hoy, y que podrá en el futuro, atribuirse a las máquinas. Al respecto, existe cierto consenso entre ciertos filósofos de la mente dedicados al estudio del concepto de la IA en considerar que la autonomía no sería una cualidad de naturaleza dicotómica, que se da o que no se da, sino una dimensión continua, por lo que no va a ser tan sencillo determinar cuándo un comportamiento realizado por una máquina va a ser autónomo y cuándo no[16]. Siguiendo el modelo de HARBERS, PEETERS y NEERINCX, existirían tres modelos de IA en la actualidad según el grado de interacción hombre-máquina: (1) *Man in the loop,* cuando la IA necesita aportes humanos a intervalos de tiempo regulares para poder llevar a cabo sus acciones; (2) *Man on the loop*, si la máquina es capaz de actuar por sí misma a partir de una programación previa, pero el humano puede intervenir interrumpiendo o modificando las acciones del robot en cualquier momento; y (3) *Man out of the loop*[17], un modelo en el que la máquina actúa de manera independiente durante ciertos períodos de tiempo y, en estos intervalos, el ser humano no tiene influencia sobre las acciones del robot. Pero la posibilidad de interacción humana en tiempo real tan sólo es una parte de lo que podríamos denominar autonomía. Por ejemplo, automatizar una respuesta ante un determinado contexto teniendo en cuenta unas variables de un modo tal que una máquina, una vez en funcionamiento, ya no se puede detener, no convertiría a la IA en autónoma. La autonomía de una IA provendría de la capacidad real de adaptar las decisiones a un contexto distinto de aquél para el que ha sido programada. Una máquina autónoma pertenecería a un cuarto tipo, digamos «*No man on the loop*», donde o bien el aprendizaje por el que se toma la decisión no hubiera devenido de una acción humana, sino de la propia máquina que ha aprendido por sí misma o de otra máquina «educadora» anterior, o bien el comportamiento de la máquina no dependiera de ese aprendizaje anterior y de las decisiones atribuidas previamente en la IA sino de algo ajeno a ello y propio del mismo sistema. Pero en la actualidad es el ser humano quien, en el proceso de aprendizaje de la máquina, determina la situación concreta a la que se ve expuesta la IA, su reacción ante una lista cerrada de estímulos, y también es el ser

[16] ANDERSON, T. L., y DONATH, M.: «Animal behavior as a paradigm for developing robot autonomy», en *Robotics and autonomous systems*, vol. 6, n.º 1-2, 1990, pp. 145-168; STEINFELD, A., FONG, T., KABER, D., LEWIS, M., SCHOLTZ, J., SCHULTZ, A., y GOODRICH, M.: «Common metrics for human-robot interaction», en *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction,* ACM, 2006, pp. 33-40; HASELAGER, W. F.: «Robotics, philosophy and the problems of autonomy», en *Pragmatics & Cognition*, vol. 13, n.º 3, 2005, pp. 515-532.

[17] HARBERS, M., PEETERS, M. M., y NEERINCX, M. A.: «Perceived autonomy of robots: Effects of appearance and context», en ALDINHAS FERREIRA, M. I., SILVA SEQUEIRA, J., TOKHI, M. O., KADAR, E., Y VIRK, G. S. (EDS.), *A World with Robots,* Springer, Cham, 2017, pp. 19-33.

humano quien no lo hace si, en la programación de la máquina, no incluye aquello que debiera haberse tomado en consideración.

2.2. La IA y su relación con el Derecho y la Justicia penal.

Los nuevos paradigmas del aprendizaje profundo y otros modelos matemáticos permiten pensar en un futuro, por tanto, en un aprendizaje propio del sistema informático, y en lo que, de un modo u otro, constituiría una IA autónoma. Eso explica que ya se hayan comenzado a plantear los retos que tales sistemas de IA supondrían para el Derecho. Ya se está planteando la necesidad de regular, cuando no directamente prohibir, el diseño de IA letales con fines militares[18]; la creación de un específico derecho robótico, como hace el proyecto RoboLaw[19]; o el control de sistemas de IA diseñados para mejorarse o autoreproducirse, como se establece en la convención de Asilomar[20], entre otras propuestas, van en la línea, parten, en mayor o menor medida, del potencial reconocimiento de identidad a las IA autónomas, y por tanto la posible incidencia, en primer lugar, en su responsabilidad, en cuanto pudieran actuar como sujetos libres, incluso como sujetos morales y por tanto capaces de ser responsables penalmente[21]; y, por otro lado, la protección de ellas mismas como bienes de especial tutela o como sujetos con derechos más o menos equivalentes a las personas[22].

---

[18] Sharkey, N: «Saying 'no!' to lethal autonomous targeting», en *Journal of Military Ethics*, vol. 9, n.º 4, 2010, pp.369-383.

[19] Palmerini, E., Bertolini, A., Battaglia, F., Koops, B. J., Carnevale, A., y Salvini, P.: «RoboLaw: Towards a European framework for robotics regulation», en *Robotics and Autonomous Systems*, vol. 86, 2016, pp.78-85.

[20] Los principios de la IA fueron desarrollados durante la convención de 2017 y pueden consultarte en la web oficial del organizador: https://futureoflife.org/ai-principles/ (Última visita en 05/09/2018).

[21] Hilgendorf, E.: «Können Roboter schuldhaft handeln?», en Beck, S. (Ed.): *Jenseits von Mensch und Maschine, Ethische und rechtliche Fragen zum Umgang mit Robotern, Künstlicher Intelligenz und Cyborgs*, Nomos, Baden-Baden, 2012, pp. 119-133, y también Hartzog, W.: «Unfair and Deceptive Robots», en *Md. L. Rev.,* vol. 74, 2014, pp. 785-829, así como Lima, D.: «Could AI Agents be held criminally liable? Artificial Intelligence and the challenges for Criminal Law», en *SCL Rev.*, vol. 69, 2017, p. 677.

[22] Véase McNally, P., y Inayatullah, S.: «The rights of robots: Technology, culture and law in the 21st century», en *Futures*, vol. 20, 1988, pp. 119-136. Sobre nada de esto me detendré por limitaciones de espacio, pero merece la pena destacar, por un lado, las reflexiones de Solum en los años 90 sobre la concesión de derechos constitucionales a máquinas autónomas rebatiendo el argumento antropocéntrico o el «*missing-something argument*», desarrollados principalmente en Solum, L. B: «Legal personhood for artificial intelligences», en *NCL Rev.*, vol. 70, 1991, pp.1231 y ss.; y, por otro, el argumento de «La habitación china» de John Searle donde rebate cualquier idea de autonomía plena en las inteligencias artificiales y la creencia de que el pensamiento se reduce a la computación (Searle, J.: «Minds, Brains and Programs», en *Behavioral and Brain Sciences*, vol. 3, 1980, pp. 417–457).

Sin embargo no sólo estamos aún lejos de lo que realmente constituiría una IA autónoma, sino que la IA que ya existe ya plantea suficientes retos en relación con el Derecho penal como para obviarlos. La IA actual, la que nos ocupa en este trabajo, aquella que consiste esencialmente en algoritmos de predicción utilizados para la realización de acciones o recomendaciones para actuar a partir de un conjunto de datos existente y de la identificación en ellos de patrones y probabilidades[23], y en la que, por tanto, todo el contexto es otorgado por los seres humanos quienes, con la información que le brindan (por acción y por omisión) y los algoritmos que crean para relacionar las variables, determinan completamente el actuar de la máquina, no requiere, a mi parecer, de ningún tipo de cambio en el sistema de atribución de responsabilidad pensado para los seres humanos como sí podría requerir en el futuro algún sistema de IA que tuviera rasgos de autonomía[24]. Pero la utilización de estos algoritmos de IA en relación con el sistema de justicia penal, por un lado, y la constatación de la posibilidad de que las máquinas, físicas o no, conformadas por IA puedan causar daños a intereses dignos de tutela penal como la vida o el orden económico, al igual que hicieron que se levantaran voces sobre la necesidad de incorporar la ética a la construcción de tales herramientas[25], nos obligan ahora a nosotros a tomar en consideración sus implicaciones penales.

Es cierto, sin embargo, que es esto último lo que más ha interesado a la dogmática penal. El hecho de que ya se hayan constatado accidentes, y otros daños distintos a los personales como los económicos[26], causados por robots, especialmente en el ámbito laboral, y también en el del tráfico rodado donde la conducción autónoma sólo empieza a desarrollar sus primeros pasos, y el que algunas de las tecnologías de IA estén comenzando a basarse en modelos de aprendizaje en los que no es sencillo definir el curso causal de la decisión tomada por la máquina, han llevado a la doctrina a revisar si el modelo de responsabilidad de la teoría del delito es adecuado para tal reto y a proponer soluciones interpretativas respecto a las diferentes situaciones de accidentes causados

---

[23] Russell, S. J., y Norvig, P.: *Artificial intelligence: a modern approach*, Pearson Education Limited, Malaysia, 2016.

[24] En tal sentido Hallevy, G.: *When Robots kill, Artificial Intelligence under Criminal Law*, Northester University Press, New England, 2013.

[25] Aunque la reflexión es anterior, véase en particular para la conducción autónoma la «denuncia» hecha por Evans, L.: «Death in Traffic: Why Are the Ethical Issues Ignored?», en *Studies in Ethics, Law, and Technology*, vol. 2, n.º 1, 2008.

[26] Lo cual también obviamos en muchas ocasiones. Respecto a la realización de daños económicos y de delitos socioeconómicos graves por medio del uso de algoritmos véase en general, Ryder, N (Ed.).: *White Collar Crime and Risk. Financial Crime, Corruption and the Financial Crisis*, Palgrave Studies in Risk, Crime and Society, Palgrave Macmillan, London, 2018, y particularmente en esa obra Baker, A.: «Market Abuse and the Risk to the Financial Markets», p. 141 y ss.

por máquinas[27]. Aunque la conclusión general a la que se llega es la de que mientras no pueda atribuirse autonomía a las entidades con IA el sistema de la teoría del delito, o mejor, los sistemas, para diferenciar el del Derecho penal anglosajón del continental, siguen siendo ante estos casos totalmente válidos para resolver los diferentes problemas causales y de atribución de responsabilidad, generalmente imprudente[28], resulta esencial monitorizar la evolución de la IA desde una perspectiva de atribución de responsabilidad para evitar llegar a situaciones en las que el aprendizaje de las máquinas no permita decir que nadie haya tomado una decisión negligente pese a que existan daños. Y no sólo eso. Muchos de los dilemas que están detrás de lo que se han denominado decisiones éticas críticas aplicables a la conducción autónoma pero, también, a muchas situaciones de riesgo en relación con el actuar de «máquinas inteligentes» que han dado lugar a importantes debates éticos [29], están íntimamente relacionadas con la atribución de responsabilidad penal, como sucede en particular con los dilemas de estado de necesidad que brillantemente ha analizado para la cuestión de la conducción autónoma Coca Vila[30], mostrando que el sistema de la teoría del delito, enraizado en la tradición filosófica y en la eterna pelea entre el consecuencialismo y el deontologicismo, puede ser un aliado para alcanzar las mejores decisiones éticas en la construcción de tales IA.

Pero, quizás porque los penalistas seguimos anclados a nuestro modo a lo físico por medio del ideal del delito de homicidio y sólo nos preocu-

---

[27] Véase, Douma, R., y Palodichuk, S. A.: «Criminal Liability Issues Created by Autonomous Vehicles», en *Santa Clara L. Rev*., vol. 52, 2012, pp. 1157-1169; Gless, S., Silverman, E., y Weigend, T.: «If Robots Cause Harm, Who Is to Blame? Self-Driving Cars and Criminal Liability», en *New Criminal Law Review: In International and Interdisciplinary Journal*, vol. 19, n.º 3, 2016, pp. 412-436; Hallevy, G.: «The Criminal Liability of Artificial Intelligence Entities-from Science Fiction to Legal Social Control», en *Akron Intellectual Property Journal*, vol. 4, n.º 2, 2010, pp. 176-177; Min-je, B.: «The Study on the Criminal Subject and Liability of AI Robots», en *International Journal of Justice and law*, vol. 2, n.º 2, 2017, pp. 15-21; Gurney, J. K.: «Driving into the unknown: Examining the crossroads of criminal law and autonomous vehicles», en *Wake Forest JL & Pol'y*, n.º 393, 2015, y el mismo, en Gurney, J. K.: «Crashing into the unknown: An examination of crash-optimization algorithms through the two lanes of ethics and law», en *Alb. L. Rev*., n.º 79, 2015, pp. 183 y ss.

[28] Véase por todos, respecto al sistema anglosajón, Hallevy, G.: *When Robots kill…*, *ob. cit.*; y al continental, Hilgendorf, E.: «Können Roboter…», *ob. cit.*

[29] Véase sobre la toma de decisiones éticas, Goodall, N.: «Ethical decision making during automated vehicle crashes», *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2424, 2014, pp. 58-65; Goodall, N. J. «Machine Ethics and Automated Vehicles», en Meyer. G., y Beiker, S. A. (Eds.): *Road Vehicle Automation*, Springer, Berlin, 2014; y también Bostrom, N., y Yudkowsky, E.: «The Ethics of Artificial Intelligence», en Frankish, K., y Ramsey, W. M. (Eds.): *Cambridge Handbook of Artificial Intelligence*, Cambridge University Press, United Kingdom, 2013.

[30] Coca Vila, I.: «Self-driving Cars in Dilemmatic Situations: An Approach Based on the Theory of Justification in Criminal Law», en *Criminal Law and Philosophy*, vol. 12, n.º 1, 2018, pp. 59-82.

pa la IA ante este tipo de resultados lesivos, o quizás por el tradicional olvido que quienes nos dedicamos al Derecho sustantivo sufrimos respecto al Derecho penitenciario y, aún más, al Derecho procesal penal, hemos prestado menos atención a los problemas que para intereses esenciales como la privacidad o para el propio funcionamiento del sistema penal conlleva la utilización en la actualidad de la IA en el sistema de justicia penal[31]. En la actualidad el uso de la IA es una realidad en la actuación policial para la prevención e investigación del delito, y está comenzando a serlo para la determinación judicial y en particular en relación con el tratamiento penitenciario, por lo que resulta esencial que su uso, en la justicia penal, se haga desde los principios propios de esta y debemos ser nosotros quienes también nos ocupemos de tal tarea.

III. Aproximación a un uso ético de la IA en el sistema de Justicia penal

3.1. Sobre el uso actual (y en el futuro) de la IA en la Justicia Criminal

Como se ha apuntado arriba, la incontestable capacidad de la IA para gestionar y procesar diferentes formas de datos con una precisión notable y para informar la toma de decisiones relacionadas con ellos ha conducido a preguntarse qué utilidad puede tener la IA en la actualidad para asistir a la justicia penal. En realidad, no cabe duda de que las utilidades potenciales pueden ser muchas también en este ámbito y, de hecho, ya están comenzando a aparecer múltiples desarrollos que en el futuro serán más y abarcarán muy probablemente un ámbito mucho mayor del que ahora imaginamos. Pero, simplificando, y a los efectos interesados de diferenciar entre los dos grandes ámbitos de utilización de la IA en EE.UU. donde más se ha desarrollado tanto el uso como la discusión académica en torno al mismo, el «policing» y el «sentencing», podríamos sistematizar los usos actuales y potenciales de la IA en relación con el sistema de justicia penal en dos grandes grupos: (1) IA para la prevención e investigación policial de la delincuencia, que denominaremos, Inteligencia Artificial Policial (IAP), e (2) IA aplicada al proceso de determinación judicial de la responsabilidad por la perpetración de un delito, que denominaríamos Inteligencia Artificial Judicial (IAJ). Aunque el interés dogmático-jurídico en España, y también en el resto del mundo, se ha centrado más en la IAJ, dado que las potenciales implicaciones para los derechos fundamentales derivadas de un mal uso de estos sistemas aparece intuitivamente como más problemáticas, creo que es necesario atender también a la IAP. No sólo por el mero hecho de que su uso en la

[31] Con la ya mencionada excepción de VALLS PRIETO, J.: *Problemas jurídico penales…, ob. cit.*

práctica es muchísimo mayor, hasta el punto de que puede anunciarse sin temor su futura generalización en relación con el ejercicio de la labor policial, sino porque sus implicaciones éticas y jurídicas no son menores, particularmente en relación con la potencial afectación a la privacidad.

3.1.1. La IA y la labor policial de prevención del crimen

En cuanto a la utilización de sistemas de IAP, y como ya se ha señalado, su popularización actual en países como EE.UU. o Inglaterra es tal que parece difícil pensar que en un futuro muy cercano los mismos no estén implantados a casi todos los niveles y ámbitos de la actuación policial. Desde una perspectiva optimista de lo que puede aportar el desarrollo tecnológico a la labor policial de prevención de la delincuencia (no sólo, pero especialmente a ella), parece claro que la IAP es la evolución natural de la aplicación de las técnicas del denominado análisis del delito. Y es que, pese al enorme conocimiento de los oficiales de policía sobre las dinámicas delincuenciales, se ha demostrado que estos no necesariamente poseen el conocimiento que les permite determinar dónde y cuándo ocurren los delitos, así como que su experiencia es más útil para investigar unos delitos que otros[32]. Y es que, en general, tratar de detectar patrones específicos de crimen y comportamientos criminales ha sido y es hoy una tarea extremadamente desafiante que exige dos cosas: el almacenamiento masivo de datos y un adecuado análisis para la extracción de inteligencia que oriente su correcta utilización.

A día de hoy somos testigos del incremento exponencial de fuentes digitalizadas de información relacionada con la actividad delictiva[33]; estadísticas oficiales de delincuencia[34]; datos obtenidos de las CCTV[35],

[32] Véase, RATCLIFFE, J. H., y MCCULLAGH, M. J.: «Chasing ghosts? Police perception of high crime areas», en *British Journal of Criminology*, vol. 41, n.º 2, 2001, pp. 330-341.

[33] MAGUIRE, M.: «Crime data and statistics», en *The Oxford handbook of criminology*, vol. 4, 2007, pp. 241-301; MAXFIELD, M. G., y BABBIE, E. R.: *Research methods for criminal justice and criminology*, Cengage Learning, USA, 2014; BACHMAN, R. D., y SCHUTT, R. K.: *Fundamentals of research in criminology and criminal justice*, Sage Publication, USA, 2016; KRISHNAMURTHY, R., y KUMAR, J. S.: «Survey of data mining techniques on crime data analysis», en *International Journal of Data Mining Techniques and Applications*, vol. 1, n.º 2, 2012, pp. 117-120.

[34] Como ejemplo paradigmático de accesibilidad libre a estadísticas oficiales de criminalidad de calidad, véase el Uniform Crime Reporting (UCR) Program del FBI en Estados Unidos (disponible en Internet en: https://ucr.fbi.gov/). En el caso de Europa tanto el European Sourcebook (ESB) of crime and criminal justice statistics (disponible en Internet en: https://wp.unil.ch/europeansourcebook/), como las Estadísticas Penales Anuales del Consejo de Europa (SPACE) (disponible en Internet en: https://www.coe.int/en/web/prison/space) permiten trabajar con datos de criminalidad de calidad con una labor mínima de preprocesamiento.

[35] Para una demostración del potencial de las imágenes de CCTV para la identificación de individuos, véase KEVAL, H. U., y SASSE, M. A.: «Can we ID from CCTV? Image

grabaciones desde las cámaras corporales policiales[36], fotografías y vídeos recogidos por drones o satélite[37], datos extraídos mediante aplicaciones móviles[38], entre otros sistemas de vigilancia; datos de cientos de millones de mensajes publicados en Internet relacionados con actos delictivos[39], ya sea en foros, chats, redes sociales, webs de compraventa, u otros espacios digitales[40]. Esto explica que los organismos e instituciones dedicados a la seguridad y a la gestión de las estadísticas de criminalidad, gobernados por la cultura del *performance* policial, estén invirtiendo una cantidad considerable de recursos en almacenar cada vez más información. Tanto es así que la obsesión por almacenar tal cantidad de información ha provocado que los analistas de delitos sean incapaces de sacarle el partido esperado[41], llegando a

---

quality in digital CCTV and face identification performance», en *Mobile Multimedia/Image Processing, Security and Applications*, 2008. DOI: 10.1117/12.774212.

[36] También denominadas *body-worn cameras*, estos dispositivos de vigilancia permiten recoger evidencias audiovisuales de determinadas acciones policiales consideradas especialmente delicadas: registros personales, persecuciones, entrevistas a testigos, o arrestos, entre otras. Aunque todavía incipiente, la investigación criminológica en *body-worn cameras* ya permite evaluar su impacto, o la percepción ciudadana sobre su uso de forma preliminar. En este sentido, véanse los trabajos de Ariel, B., Farrar, W. A., y Sutherland, A.: «The effect of police body-worn cameras on use of force and citizens' complaints against the police: A randomized controlled trial», en *Journal of Quantitative Criminology*, vol. 31, n.º 5, 2015, pp. 509-535; Ariel, B., Sutherland, A., Henstock, F., Young, J., Drover, P., Sykes, J., Megicks, S., y Henderson, R.: «Report: increases in police use of force in the presence of body-worn cameras are driven by officer discretion: a protocol-based subgroup analysis of ten randomized experiments», en *Journal of Experimental Criminology*, vol. 12, n.º 3, 2016, pp. 453-463; Jennings, W. G., Fridell, L. A., y Lynch, M. D.: «Cops and cameras: Officer perceptions of the use of body-worn cameras in law enforcement», en *Journal of Criminal Justice*, vol. 42, n.º 6, 2014, pp. 549-556.

[37] Para una revisión de los aspectos éticos y legales derivados del uso de datos recogidos por drones y sus diversas aplicaciones en Big Data, véase Finn, R., y Donovan, A.: «Big Data, Drone Data: Privacy and Ethical Impacts of the Intersection Between Big Data and Civil Drone Deployments», en Custers, B. (Ed.) *The Future of Drone Use,* TMC Asser Press, La Haya, 2016, pp. 47-67.

[38] Para un ejemplo del tipo de datos que pueden revelar las aplicaciones móviles, véase Hern, A.: «Fitness tracking app Strava gives away location of secret US army bases», en *The Guardian* [en línea], 28 de enero 2018. Disponible en: https://www.theguardian.com/world/2018/jan/28/fitness-tracking-app-gives-away-location-of-secret-us-army-bases (Última visita el 05/09/2018)

[39] Véase Reaves, B.: *Local Police Departments, 2013: Equipment and technology,* Bureau of Justice Statistics, Washington, DC, 2015.

[40] Sobre la utilización de datos procedentes de Internet para el análisis delictivo véase McCue, C.: *Data mining and predictive analysis: intelligence gathering and crime analysis,* 2.ª edición, Butterworth-Heinemann, Oxford, UK, 2014.

[41] En este sentido, véase Ratcliffe, J. H.: *Intelligence-led Policing*, Willian Publishing, Portland, OR, 2008, sobre la complejidad de la gestión de la información debido a la *performance culture* que ha incrementado considerablemente la burocracia, la monitorización de la contabilidad, o la fiscalización interna a la que están sometidos los trabajadores. A esto hay que añadir las diferentes formas de registrar los datos según el departamento en relación con su formato o sus variables de interés, una circunstancia que dificulta su tratamiento y estandarización.

invertir incontables horas revisando datos sobre diversas formas de delincuencia para determinar si un delito encaja en un patrón conocido o, por el contrario, forma parte de un nuevo patrón desconocido. Tal ineficiencia ha evidenciado la necesidad de incorporar estas IAP que, frente a los modelos analíticos tradicionales[42], están empezando a ofrecer grandes ventajas reduciendo las tareas de análisis y supervisión.

Estas herramientas, basadas particularmente en el conocimiento teórico de la denominada criminología ambiental[43], se nutren de algoritmos matemáticos para la selección y ubicación de los recursos policiales que deben ser dedicados a la prevención de la criminalidad y que, junto a la aplicación de las técnicas geoestadísticas del análisis delictivo, constituyen una herramienta de gran potencial. Y el ejemplo paradigmático es lo que ha venido a denominarse «*predictive policing*» o también *PredPol*[44]. En concreto, este conjunto de IAP, de modo general, se basan en la aplicación de técnicas cuantitativas para identificar objetivos de interés policial con el propósito de reducir el riesgo delictivo mediante la prevención de delitos futuros o la resolución de delitos pasados[45]. Así, este enfoque analítico trata de aprovechar el poder de la información, las tecnologías de georreferenciación, los avances de la criminología ambiental y los modelos de intervención policiales basados en evidencias para reducir la delincuencia y mejorar la seguridad pública, y con ello permitir a las agencias de seguridad pasar de la reacción a los delitos a predecir qué y dónde es más probable que ocurra algo y, como resultado, poder

[42] Witten, I. H., Frank, E., Hall, M. A., y Pal, C. J.: *Data Mining: Practical machine learning tools and techniques*, Morgan Kaufmann, Cambridge, MA, 2016.

[43] Principalmente en las estrategias de Policía Orientada a la Solución de Problemas y de Prevención Situacional del Crimen. En este sentido, véase respectivamente Goldstein, H. «Improving policing: A problem-oriented approach», en *Crime & delinquency*, vol. 25, n.º 2, 1979, pp. 236-258.; CLARKE, R. V. G. «Situational crime prevention: Theory and practice», en *British Journal of Criminology*, vol. 20, 1980, pp. 136-147. En este sentido, el Centro para la Policía Orientada a la Solución de Problemas ha recogido y sistematizado en su plataforma online a lo largo de los años algunos de estos recursos para apoyar a policías, investigadores y universidades dedicados al fomentar el progreso de la POP (véase en Internet en: www.popcenter.org).

[44] Véase Perry, W. L., McInnis, B., Price, C. C., Smith, S. C., y Hollywood, J. S.: *Predictive Policing. The role of crime forecasting in Law Enforcement operations,* CA: RAND Corporation, Santa Monica, 2013. Para otras introducciones de interés a este concepto, véase Ferguson, A. G.: «Policing predictive policing», en *Wash. UL Rev.*, vol. 94, 2016, pp. 1109 y ss.; Lum, K., y Isaac, W.: «To predict and serve?» en *Significance*, vol. 13, n.º 5, 2016, pp. 14-19.

[45] Perry y colaboradores agrupan las principales técnicas de *data mining* para realizar análisis predictivos en dos grandes categorías: (1) técnicas de clasificación que pretenden establecer reglas y condiciones para poder etiquetar eventos; y (2) técnicas de clústeres para agrupar los datos en conjuntos con atributos similares. Para profundizar en estos aspectos, véase Perry, W. L., McInnis, B., Price, C. C., Smith, S. C., y Hollywood, J. S.: *Predictive Policing…, ob. cit.* pp., 33-41.

desplegar los recursos necesarios[46]. Nunca antes, o al menos no en la misma medida, la previsión del delito había estado orientada hacia enfoques tan empíricos y basados en datos cuya utilidad se manifiesta tanto a la hora de diseñar operativos policiales para responder ante problemas concretos, como para reforzar la investigación de agreso-

[46] En términos generales, está la definición que ofrece el *National Institute of Justice* de Estados Unidos del *predictive policing*, cfr. http://www.predpol.com/category/predictive-policing/. Desde la toma en consideración de este nuevo paradigma policial, existen numerosos ejemplos de IA aplicada a la prevención e investigación policial de la criminalidad. En este sentido, véase, principalmente, Perry, W. L., McInnis, B., Price, C. C., Smith, S. C., y Hollywood, J. S.: *Predictive Policing…*, *ob. cit.*, pp. 1-2; y el trabajo de Ratcliffe, J. H.: *Intelligence-led…*, *ob. cit.*, sobre la adopción de modelos policiales guiados por inteligencia extraída, entre otras fuentes, de distintas técnicas de análisis delictivo. Adicionalmente, véanse las contribuciones presentadas en los *Predictive Policing Symposiums* organizados por el *National Institute of Justice* de los Estados Unidos (US Department of Justice, *Predictive Policing* Symposiums, 2009-2010. Disponibles en Internet en: https://www.ncjrs.gov/pdffiles1/nij/242222and248891.pdf). Desde la perspectiva de la prevención de la delincuencia en entornos físicos, y especialmente en áreas urbanas, son destacables los algoritmos para el geoposicionamiento como los desarrollados recientemente por Caplan, J. M., Kennedy, L. W., y Miller, J.: «Risk terrain modeling: brokering criminological theory and GIS methods for crime forecasting», en *Justice Quarterly*, vol. 28, n.º 2, 2011, pp.360-381 o Kennedy, L. W., Caplan, J. M., y Piza, E.: «Risk clusters, hotspots, and spatial intelligence: risk terrain modeling as an algorithm for police resource allocation strategies», en *Journal of Quantitative Criminology*, vol. 27, n.º 3, 2011, pp. 339-362, quienes explican cómo el RTM es, a todos los efectos, un método de diagnóstico que permite realizar pronósticos muy precisos mediante el análisis de los factores ambientales que generan y atraen el delito. Siguiendo una línea similar, el proyecto de *Riskment* del Centro CRÍMINA para el estudio y prevención de la delincuencia ha desarrollado un algoritmo para identificar los segmentos de vía de las provincias de Alicante y Cádiz en los que es más probable, a partir de una serie de factores, que haya una conducción influenciada o un accidente relacionado con ella. Para lograr este objetivo, se han analizado todos los datos pasados de accidentalidad en esos lugares, identificado patrones, asociado variables a partir de marcos teóricos y valorado dónde es más probable que pasen los accidentes según las citadas variables. Por su parte, las IA para el geoposicionamiento del delito han ido más allá en la actualidad y han introducido otros sensores para recopilar información de interés. Sírvanos como ejemplo que la compañía china *Hikvision* ha desarrollado un nuevo modelo de cámara de vigilancia que realiza funciones de investigación que incluyen la lectura de placas de matrícula, la búsqueda de bultos sospechosos en áreas masificadas, o la identificación mediante reconocimiento facial de individuos sospechosos. En este caso, la compañía afirma alcanzar precisiones con su sistema de análisis visual del 99%, que en la actualidad está siendo aplicado, entre otros ámbitos, a la prevención frente actividades radicales islámicas (véase en Internet en: https://www.ft.com/content/c610c88a-8a57-11e8-bf9e-8771d5404543). O también el algoritmo para el geoposicionamiento de sonidos de armas de fuego que aprovecha la infraestructura de algunas de las ciudades inteligentes o *smart cities* estadounidenses tras la implantación de un sistema de triangulación de sonidos que se sirve del *machine learning* para, gracias a un conjunto de sensores, poder ubicar geográficamente el lugar desde donde se ha disparado un arma de fuego. Los desarrolladores de la tecnología *ShotSpotter* afirman que su sistema puede geoposicionar este tipo de sonidos con una precisión de 10 pies (véase en Internet en: https://baltimore.cbslocal.com/2018/06/01/shotspotter-baltimore-police/).

res potenciales con un elevado nivel de peligrosidad. Respecto a los primeros, las intervenciones policiales guiadas por IAP han mostrado reiteradamente su utilidad a la hora de reducir la delincuencia en aquellos lugares identificados como de alto riesgo donde se generan concentraciones desproporcionadas o *hot spots* de criminalidad[47]. En cuanto a lo segundo, la incorporación de información sobre patrones comportamentales de ciertos individuos en los algoritmos predictivos permite estimar quién es más probable que cometa un delito, priorizar un sujeto de una larga lista de sospechosos, o incluso asignar uno u otro programa de reinserción a un agresor en función de sus características personales[48].

En relación con este segundo grupo de herramientas, además de la predominante intervención policial basada en lugares, se están comenzando a implantar modelos predictivos que permiten hacer prevención policial en individuos. Quizá el ejemplo más conocido sea el algoritmo de Rossmo, cuya formulación permite estimar el área geográfica donde, con mayor probabilidad, reside un presunto agresor serial en función de la ubicación de los delitos que previamente se le atribuyen[49]. Gracias a tal estimación, es posible priorizar sospechosos de una larga lista y concentrar los recursos policiales en determinados individuos que encajan con la geografía del crimen. Otros instrumentos, como *The Lethality Screen*, se están utilizando de forma conjunta con herramientas de valoración del riesgo de violencia doméstica para proporcionar información útil a los equipos de intervención temprana. Este sistema ha mostrado tanto un alto valor predictivo para los casos en los que es improbable que se ejecuten actos de violencia letal o severa, como un elevado grado de concordancia con la percepción del riesgo de las víctimas[50]. Siguiendo una línea similar, en España se ha desarrollado el sistema de seguimiento integral VioGén para, a través de un algoritmo que analiza la información actuarial disponible en el protocolo de valoración policial del riesgo (VPR)[51], tratar de identificar aquellos casos en los que la victimización por violencia doméstica es más probable.

---

[47] Braga, A. A., Weisburd, D.: *Policing problem places: Crime hot spots and effective prevention*. Oxford University Press on Demand, 2010.

[48] Sun, K. *Correctional counseling: A cognitive growth perspective*. Jones & Bartlett Learning, 2008.

[49] Rossmo, D. K.: *Geographic Profiling: Target Patterns of Serial Murders*. 1995. Tesis doctoral. Simon Fraser University.

[50] Messing, J. T., Campbell, J., Sullivan Wilson, J., Brown, S., & Patchell, B.: «The lethality screen: the predictive validity of an intimate partner violence risk assessment for use by first responders», en *Journal of interpersonal violence*, vol. 32, n.º 2, 2017, p. 205-226.

[51] López-Ossorio, J. J., González-Álvarez, J. L., & Andrés-Pueyo, A.: «Eficacia predictiva de la valoración policial del riesgo de la violencia de género», en *Psychosocial Intervention*, vol. 25, n.º 1, 2016, pp. 1-7. En su trabajo de evaluación, estos autores encontraron que el protocolo tenía una capacidad predictiva notable y un diseño apropiado.

Siguiendo las directrices marcadas por el correspondiente Protocolo, los funcionarios de la policía especializados en el tratamiento de este tipo de violencia pueden proporcionar a la víctima las medidas de protección que más se ajusten a cada situación en función del riesgo que haya determinado el VPR[52]. Respecto al agresor, se procederá a la incautación de armas y/o instrumentos peligrosos que se hallen en su poder o en el domicilio familiar y, cuando el riesgo sea bastante, se deberá proceder a su detención y puesta a disposición judicial[53]. Teniendo en cuenta la potencial utilidad de los resultados obtenidos tras las primeras aplicaciones de estos sistemas de IAP y la innegable capacidad preventiva derivada de su utilización, y pese a que estas herramientas aún no pueden considerarse IAP en sentido estricto, parece lógico pensar que su perfeccionamiento pronto contribuirá a su progresiva implantación.

Por otro lado, y más allá de la prevención e investigación de la delincuencia tradicional o en entornos físicos, los avances de las Tecnologías de la Información y la Comunicación han generado nuevas oportunidades delictivas en el ciberespacio que aprovechan los delincuentes para cometer numerosos delitos[54]. Las agencias de seguridad, que necesitan responder ante las nuevas demandas de seguridad, empiezan a poner cada vez más de relieve sus necesidades en materia de predicción y prevención de la ciberdelincuencia. Es por ello que, aunque tradicionalmente el foco de atención se ha centrado en la detección de vulnerabilidades en infraestructuras digitales que puedan dar lugar a intrusiones y otras ciberamenazas, durante la última década se han venido desarrollando IAP cada vez más sofisticadas, flexibles, adaptables y robustas, capaces de detectar y prevenir una amplia variedad de amenazas y tomar decisiones inteligentes en

---

[52] Consejo General del Poder Judicial: *Protocolo de actuación de las Fuerzas y Cuerpos de Seguridad y de coordinación con los órganos judiciales para la protección de las víctimas de violencia doméstica y de género*, 28 de junio de 2005 (disponible en Internet en: http://www.violenciagenero.igualdad.mpr.gob.es/profesionalesInvestigacion/seguridad/protocolos/pdf/Protocolo_Actuacion_Fuerzas_Cuerpos_Seguridad_Coordinacion_Organos_Judiciales.pdf). Las medidas específicas que recoge el Protocolo para estos casos son: (1) protección personal a la víctima por un periodo de tiempo establecido en función del riesgo calculado; (2) formación respecto a la adopción medidas de autoprotección; (3) provisión de información comprensible sobre el contenido, tramitación y efectos de la orden de protección y seguridad previstas en la Ley, así como de otros servicios de asistencia; y (4) en el caso de víctimas extranjeras, información sobre su derecho a regular su situación por razones humanitarias.

[53] Véase, Consejo General del Poder Judicial: *Protocolo de…*, *ob. cit.* p. 7.

[54] Miró Llinares, F.: «La oportunidad criminal en el ciberespacio: Aplicación y desarrollo de la teoría de las actividades cotidianas para la prevención del cibercrimen», en *Revista Electrónica de Ciencia Penal y* Criminología, n.º 13, 2011. Para una revisión de las diferentes formas de cibercriminalidad, véase Miró Llinares, F.: *El cibercrimen. Fenomenología y criminología de la delincuencia en el ciberespacio,* Marcial Pons, Madrid, 2012.

tiempo real[55]. Así, frente al interés por la predicción y prevención de la ciberdelincuencia más técnica, existe un interés creciente por el denominado Factor Humano del cibercrimen y la ciberseguridad[56], que enfatiza el estudio de la motivación de los ciberdelincuentes, las redes de organizaciones criminales, las actividades cotidianas *online* de las potenciales víctimas, o las distintas formas de atajar este tipo de cibercriminalidad más social. En este sentido, también empieza a darse una preocupación policial transfronteriza por el papel que ocupan las redes sociales *online* en los procesos de radicalización y en la difusión de contenidos de esta naturaleza[57], y en consecuencia por la necesidad de desarrollar herramientas aptas para realizar análisis automatizado de grandes cantidades de datos procedentes de estas plataformas digitales que sean útiles en la investigación, predicción y prevención de estas otras formas de criminalidad[58]. En ambos casos,

---

[55] En especial, son numerosos métodos de computación bioinspirados de IA (como redes neuronales, *Intelligent Agent Application, Artificial Inmune System Application, Genetic Algorithm y las Fuzzy Sets Applications*) los que están jugando un papel cada vez más importante en la detección y prevención de diferentes ciberdelitos de naturaleza económica. En esta línea, cabe mencionar la revisión de la literatura que hacen Dilek, S., Çakir, H., y Aydin, M.: «Applications of artificial intelligence techniques to combating cyber crimes: A review», 2015 (disponible en: *arXiv preprint arXiv:1502.03552)*, donde presenta los avances logrados hasta la fecha en la aplicación de las técnicas de inteligencia artificial en la lucha contra diferentes formas de cibercriminalidad y explica cómo estas técnicas pueden ser una herramienta eficaz para la detección y prevención de los ataques a infraestructuras digitales.

[56] Véase el trabajo coordinado por el *Netherlands Institute for the Study of Crime and Law Enforcement* Leukfeldt, R.: *Research Agenda: The Human Factor in Cybercrime and Cybersecurity*, Eleven International Publishing, The Hague, 2017.

[57] En este sentido, véase el extenso trabajo sobre *data science* y ciberseguridad en redes sociales del grupo de investigación de la Universidad de Cardiff liderado por el Prof. Pete Burnap y Matthew L. Williams. Por ejemplo, Burnap, P., Williams, M. L., Sloan, L., Rana, O., Housley, W., Edwards, A., Knight, V., Procter, R., y Voss, A.: «Tweeting the terror: modelling the social media reaction to the Woolwich terrorist attack», en *Social Network Analysis and Mining*, vol. 4, n.º 1, 2014, p. 206 y ss.; o Burnap, P., y Williams, M. L.: «Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making», en *Policy & Internet*, vol. 7, n.º 2, 2015, pp. 223-242.

[58] Tradicionalmente, los esfuerzos dedicados al análisis de redes sociales para la detección de contenidos considerados ilícitos se han circunscrito a la búsqueda de patrones tanto sintácticos como semánticos y, para ello, se han empleado distintas técnicas de análisis de contenido que han puesto el acento en las palabras como unidad de análisis, ya sea mediante la búsqueda sistemática de palabras clave (Cfr. Décary-Hétu, D., y Morselli, C.: «Gang presence in social network sites», en *International Journal of Cyber Criminology*, vol. 5, n.º 2, 2011, pp. 876-890), a través de sistemas de análisis de sentimiento (Cfr. Cheong, M., y Lee, V. C.: «A microblogging-based approach to terrorism informatics: Exploration and chronicling civilian sentiment and response to terrorism events via Twitter», en *Information Systems Frontiers*, vol. 13, n.º 1, 2011, pp. 45-59.*)* o polarización del discurso político (Cfr. Conover, M., Ratkiewicz, J., Francisco, M. R., Gonçalves, B., Menczer, F., y Flammini, A.: «Political polarization on twitter», en *ICWSM*, n.º 133, 2011, pp. 89-96), entre otros métodos, para estudiar las actitudes de los emisores a la hora de publicar estos mensajes. En esta línea, ha cobrado especial relevancia el diseño de herramientas infor-

estas IAP suelen disponer de arquitecturas basadas en la inclusión de otras fuentes de datos exclusiva de este nuevo entorno digital: los metadatos. Más concretamente, integran en sus procesos analíticos variables que, como el medio de creación de los datos, finalidad de los datos, hora y fecha de creación, creador o autor de los datos, ubicación en una red informática donde se crearon los datos, normas utilizadas, tamaño de archivo, calidad de los datos, fuente de los datos, proceso utilizado para crear los datos, información externa vinculada con los datos, entre otros muchos, han mostrado una gran capacidad para la predicción de ciberdelitos[59].

3.1.2. La IA en el ámbito judicial penal

Aún más interés está suscitando a nivel académico y social la utilización de tecnologías de Inteligencia Artificial en el propio sistema de Justicia penal, tanto por las enormes implicaciones que su uso puede tener en relación con una potencial mejora del sistema de justicia como por los múltiples riesgos que su uso puede suponer para derechos y garantías fundamentales de nuestro sistema penal. Si bien, y frente a la

---

máticas que automaticen los procesos de detección de contenido radical en poco tiempo y con elevados niveles de precisión, como apoyo a las arduas tareas de vigilancia destinadas a la detección temprana de contenido radical en Internet, como los análisis avanzados sobre crimen que provee Wynyard (Wynyard Group), revelando aquella información procesable que se encuentra oculta en los datos; el Observatorio Colaborativo Online de Redes Sociales (COSMOS), que ofrece un servicio integrado y escalable para analizar redes sociales según determinadas demandas (Burnap, P., Rana, O., Williams, M., Housley, W., Edwards, A., Morgan, J., Sloan, L., y Conejero, J.: «COSMOS: Towards an integrated and scalable service for analysing social media on demand», en *International Journal of Parallel, Emergent and Distributed Systems*, vol. 30, n.º 2, 2015, pp. 80-100); o el *software* para análisis de la delincuencia que proporciona Watson (IBM), que permite explorar la información disponible en las redes sociales con su funcionalidad cognitiva. Con la misma finalidad, cabe señalar el desarrollo de la *Cyberspace Detection Tool* o IA de detección y clasificación de mensajes de radicalización en Internet realizada por CRÍMINA para el proyecto europeo PERICLES. Para el desarrollo de esta herramienta ha sido necesaria la compilación de millones de tuits que fuimos recogiendo tras los atentados terrorista de Charlie Hebdo, Paris Bataclan, Bruselas, Orlando, Barcelona, etc., se ha construido una herramienta para la detección y clasificación de mensajes radicales mediante la combinación de herramientas de análisis de texto y técnicas de árboles de decisión, que aprenden a clasificar contenidos de forma más eficaz conforme analizan muestras nuevas. A día de hoy, estamos alcanzado precisiones que rondan el 80% a la hora de identificar discurso radical en Twitter (veáse en Internet en: http://project-pericles.eu/).

[59] Véase, Esteve, M., Miró Llinares, F., y Rabasa, A.: «Classification of tweets with a mixed method based on pragmatic content and meta-information», en *Complex System Studies*, vol. 13, n.º 1, 2018, pp. 60-70; Schmidt, A., y Wiegand, M.: «A survey on hate speech detection using natural language processing», en *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media,* abril 3-7, 2017, pp. 1-10.; Waseem, Z., y Hovy, D.: «Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter», en *Proceedings of NAACL-HLT*, junio 12-17, 2016, pp. 88-93.

popularización del uso de los algoritmos de «policía predictiva», aún no puede decirse que la Inteligencia Artificial Judicial (IAJ) se haya generalizado, la utilización de sistemas de IA en distintas fases del proceso judicial penal por los diferentes operadores jurídicos empieza a no ser extraordinaria y todo parece indicar que se hará aún más habitual en el futuro. Como recientemente ha puesto de manifiesto NIEVA FENOLL, son múltiples los posibles usos de la IA en la justicia penal[60]: tanto en materia de procedimiento, particularmente en relación con la sistematización y análisis de la ingente documentación que suele haber en los procesos y con la finalidad de mejorar la eficiencia (especialmente temporal) en la tramitación, como de prueba y de argumentación por medio de herramientas «que pueden ayudar al juez a valorar la prueba, o al menos a ordenar su razonamiento» como STEVIE, EMBRACE, ECHO, PEIRCE-IGTT[61], entre otras muchas herramientas que ya están siendo usadas[62]. Pese a la potencial utilidad para la mejora de la justicia de todas estas formas de IAJ, obviamente el máximo interés lo tienen aquellas que informan a los tribunales para la toma de decisiones en el ámbito penal, en particular las que han sido diseñadas con el

[60] Véase, con profundidad y detalle de cada una de las utilidades, NIEVA FENOLL, J.: Inteligencia artificial y proceso …, *ob. cit*., pp. 23 y ss.

[61] Todas ellas magníficamente explicadas por NISSAN, E.: «Digital technologies and artificial intelligence's present and foreseeable impact on lawyering, judging, policing and law enforcement», en *AI & Society,* vol. 32, n.º 3, 2017, pp. 441y ss., y por NIEVA FENOLL, J.: Inteligencia artificial y proceso …, *ob. cit*.

[62] Entrarían, pues, aquellas técnicas de procesamiento y análisis de datos que de hecho ya están sirviendo para agilizar la propia gestión y toma de decisiones judiciales. Entre ellos, cabe destacar *Prometea*, *software* desarrollado en Argentina que actualmente se está utilizando en el Tribunal Superior de Justicia porteño para predecir la solución de expedientes jurídicos que los creadores califican como «simples», refiriéndose con este calificativo a aquellos casos que van desde recursos de amparo por supuesta violación del derecho a la vivienda, a causas de empleo público o cuestiones meramente procesales, siempre que haya precedentes muy similares y la solución jurídica sólo sea una. Esta suerte de asistente judicial no actúa solo, sino que requiere que una persona le dé indicaciones acerca de cierta información relevante sobre el caso, para posteriormente analizar todas las sentencias de primera y segunda instancia de la Ciudad Autónoma de Buenos Aires, y con ello sugerir un modelo de respuesta, un dictamen. Obviamente, una vez que *Prometea* emite un dictamen, el documento que genera pasa por una revisión humana sin excepción (Cfr. CORVALÁN, J. G.: «Artificial Intelligence, Threats, Challenges and Opportunities-Prometea, the First Predictive Artificial Intelligence at the Service of Justice is Argentinian», en *Revue Internationale de droit des données et du numérique*, vol. 4, 2018, pp. 17-36.). Con esta misma orientación, pero sin un implante en la praxis judicial, se han creado otras herramientas por investigadores, como la que se creó en Reino Unido por Barot y Carter en 2008, denominada JAES (Judicial Advisory Expert System), para la implementación de un algoritmo Q-learning para optimizar el tiempo de toma de decisiones en el sistema judicial (Cfr. BRENNAN, T., y DIETERICH, W.: «Correctional Offender Management Profiles for Alternative Sanctions (COMPAS)», en SINGH, J. P., KRONER, D. G., WORMITH, J. S., DESMARAIS, S. L., y HAMILTON, Z. (EDS.), *Handbook of Recidivism Risk/Needs Assessment Tools*, John Wiley & Sons, 2018, pp. 49 y ss.; GIFFORD, R.: «Legal technology: Criminal justice algorithms: Ai in the courtroom», en *The Proctor,* vol. 38, n.º 1, 2018, pp. 32 y ss.).

propósito de evaluar automáticamente los factores de riesgo supuestamente predictivos del actuar delictivo futuro[63].

Los sistemas de IAJ de valoración del riesgo (IAJVR, a partir de ahora) están basados en la automatización de las herramientas de valoración del riesgo[64], y son potencialmente aplicables a los procesos de toma de decisiones relacionados con la valoración del riesgo de aquellos individuos que se ven envueltos en el proceso judicial, tanto durante él, como puede ser para la aplicación de medidas cautelares, como tras la condena. Es decir, que se trata de herramientas que informan la toma de decisiones tan importantes como la concreción del régimen penitenciario, la concesión de la libertad provisional y la libertad condicional, la reubicación del reo en uno u otro régimen penitenciario, entre otras[65]. Como sucedía con el primer conjunto de herramientas, el

---

[63] Más concretamente, la literatura científica ha establecido que las variables de predicción más fuertes de la reincidencia han sido las necesidades criminógenas, la historia criminal y de comportamientos antisociales, los logros sociales, la edad, el género, la raza y los factores familiares. Otros predictores menos robustos han sido el funcionamiento intelectual, los factores de angustia personal y el estatus socioeconómico en la familia de origen. A día de hoy, las herramientas actuariales para la evaluación del riesgo de reincidencia son instrumentos fundamentales para la toma de decisiones con consecuencias jurídicas. Para un metaanálisis sobre los predictores más importantes de la reincidencia en edad adulta, véase: Gendreau, P., Little, T., y Goggin, C.: «A meta-analysis of the predictors of adult offender recidivism: What works!», en *Criminology*, vol. 34, n.º 4, 1996, pp. 575-607. Para el caso de reincidentes sexuales, véase el siguiente metaanálisis: Hanson, R. K., y Bussiere, M. T.: «Predicting relapse: a meta-analysis of sexual offender recidivism studies», en *Journal of consulting and clinical psychology*, vol. 66, n.º 2, 1998, pp. 348-362.

[64] Algunas de estas herramientas son: el HCR-20, que permite realizar un juicio profesional estructurado para la valoración y gestión del riesgo de violencia, Douglas, K. S., Hart, S. D., Webster, C. D., y Belfrage, H.: *Assesing risk for violence.* Version 3, Mental Health, Law and Policy Institute, Simon Fraser University, Burnaby, BC, 2013; el SAVRY, un juicio estructurado para la valoración y gestión del riesgo de violencia juvenil, Borum, R., Bartel, P., y Forth, A.: «Structured Assessment of Violence Risk in Youth», en Grisso, T., Vincent, G., y Seagrave, D. (Eds.), *Mental Health Screening and Assessment in Juvenile Justice*, The Guilford Press, New York, 2005; el PCL-R, una escala revisada para evaluar la psicopatía, Hare, R. D.: *The Hare Psychopaty Checklist-Revised*, 2.ª Edición, Multi-Health Systems, Toronto, ON, 2003, también en su versión para jóvenes, el PCL:YV, Forth, A. E., Kosson, D. S., y Hare, R. D.: *Hare psychopaty checklist: Youth versión,* Multi-Health Systems, Toronto, ON, 2003; el SVR-20 para un juicio estructurado del riesgo de violencia de carácter sexual, Boer, D. P., Hart, S., Kropp, P. R., y Webster, C. D.: *The SVR-20 Guide for assessment of sexual risk violence,* Mental Health, Law and Policy Institute, Simon Fraser University, Vancouver, 1997; o la SARA para el caso concreto de la violencia contra la pareja, Kropp, P. R., Hart, S., Webster, C. D., y Eaves, D.: *Manual for the Spousal Assault Risk Assesment Guide,* 2.ª edición, British Columbia Institute on Family Violence, Vancouver, 1995. Para una revisión en castellano de los instrumentos disponibles en el campo de la valoración del riesgo de violencia y su aplicación, véase también Andrés-Pueyo, A., y Echeburúa, E.: «Valoración del riesgo de violencia: instrumentos disponibles e indicadores de aplicación», en *Psicothema*, vol. 22, n.º 3, 2010, pp. 403-409.

[65] Rizer, A., y Watney, C.: *Artificial Intelligence Can Make Our Jail System More Efficient, Equitable and Just,* 2018. Disponible en Internet en: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3129576

funcionamiento de este tipo de IA es el mismo: a partir de un set de datos se busca el establecimiento de patrones y relaciones entre variables para establecer una predicción respecto al comportamiento futuro. Sin embargo la diferencia es que en lugar de usar datos georreferenciados o metadatos, las herramientas actuariales automatizadas se apoyan en datos personales relacionados con factores individuales, sociales y ambientales, y en la propia configuración de las herramientas de valoración del riesgo de violencia y reincidencia basadas en metodologías cualitativas y cuantitativas de diferente tipo pero que van más allá del tradicional juicio psicológico clínico al tratarse de estudios analíticos sistematizados que se están utilizando desde hace décadas en relación con algunos delitos graves aunque empiezan a generalizarse para otros muchos[66]. En todo caso, y después profundizaré sobre esto, no deben confundirse las herramientas actuariales de valoración del riesgo con las herramientas de IAJVR que ya existen y que se desarrollarán en el futuro. Eso sí, la diferencia es sutil, y no sólo porque la segunda implica la primera (aunque no al revés), sino porque casi todas las herramientas de valoración del riesgo en la actualidad tienen algo automatizado si bien para que sean IA la intervención de la máquina deberá ser esencial.

El ejemplo más conocido de IAJVR es COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*), empleada en Estados de Estados Unidos, para evaluar el riesgo de reincidencia de los delincuentes con un triple objetivo: a) construir una herramienta que supere los sesgos existentes en los tribunales juzgadores y que, basándose en evidencias, sea capaz de mejorar la toma de decisiones; b) ayudar al personal correccional a asignar a los reclusos correctos a los programas correctos en el momento adecuado, basándose en evaluaciones individuales de riesgos y necesidades, incluyendo la provisión de programas de rehabilitación para los prisioneros de mayor riesgo de reincidencia y presos en libertad condicional, y proporcionar otros tipos de programas para los prisioneros de bajo riesgo de reincidencia y presos en libertad condicional, y c) ayudar a reducir la probabilidad de que el recluso reincida cuando regrese a la sociedad[67]. Creada por la empresa Northpoint, Compas es una IA que, por medio de un algoritmo no público, por estar protegido por la ley de propiedad intelectual, combina en su análisis datos de diferente naturaleza[68] para evaluar el

[66] Véase particularmente respecto a los factores que suelen utilizar estas herramientas ANDRÉS-PUEYO, A., y ECHEBURÚA, E.: «Valoración del riesgo de violencia: instrumentos disponibles e indicadores de aplicación», en *Psicothema*, vol. 22, n.º 3, 2010.

[67] BRENNAN, T., y DIETERICH, W.: «Correctional Offender…», *ob. cit.*, pp. 49 y ss; GIFFORD, R.: «Legal technology: Criminal justice algorithms: Ai in the courtroom», en *The Proctor,* vol. 38, n.º 1, 2018, pp. 32 y ss.

[68] Describe NIEVA FENOLL los ítems como «variopintos», quien añade más adelante que muchos de ellos no sólo no se pueden relacionar con propensión alguna al delito

riesgo de que una persona cometa un delito, tanto en el caso de que esté siendo enjuiciada y se plantee la prisión provisional, como en el de que ya haya sido sancionada y tenga que determinarse el castigo concreto que se le aplicará y su forma de ejecución. Pese a que lleva varios años en funcionamiento ha sido recientemente cuando la misma se ha hecho más visible debido a una resolución judicial y a un reportaje periodístico. El reportaje lo publicó la web ProPublica, con el sugerente título «El sesgo de la máquina: Existe un *software* que se está usando en todo el país para la predicción de futuros criminales. Y está sesgado contra los negros», y en él se analizaba la fiabilidad de la predicción realizada por COMPAS y se ponía en duda la equidad de un sistema que parecía discriminar significativamente a aquel grupo de población[69]. En cuanto a la sentencia, es la del caso State vs. Loomis del Tribunal Supremo de Wisconsin[70], que resolvía la apelación presentada por Eric Loomis por violación de derechos constitucionales, como el derecho a un debido proceso y el de igualdad por la utilización de tal herramienta para la determinación de las altas posibilidades de reincidencia que incidió en la concreción de la pena de prisión. Aunque sobre esto volveré después, la resolución no atendió a las demandas del apelante y dio por buena la utilización de este tipo de herramientas para la valoración del riesgo de reincidencia siempre que

---

sino que son «directamente clasistas», NIEVA FENOLL, J.: Inteligencia artificial y proceso …, *ob. cit.*, pp. 68 y 69. Sin entrar en estas consideraciones, lo cierto es que aunque gran parte de los indicadores utilizados por COMPAS son los tradicionalmente usados en las herramientas actuariales de valoración del riesgo de reincidencia hay otros en los que no es fácil identificar el marco teórico que hay detrás, y la falta de publicidad del algoritmo agrava las dudas sobre su sentido.

[69] Véase sobre esto ANGWIN, J.; LARSON, J.; MATTU, S.; KIRCHNER, L.: «Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks», en *ProPublica,* [en línea], 23 de mayo de 2016. Disponible en: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (Última visita el 21/10/2018). Hubo una respuesta de NORTHPOINTE.INC: «Response to ProPublica: demonstrating accuracy equity and predictive parity». Disponible en http://www.equivant.com/blog/response-to-propublica-demonstrating-accuracy-equity-and-predictive-parity (Última visita el 21/10/2018); sin embargo más interesante es el debate académico que surgió después, con decenas y decenas de artículos en sólo dos años, en el que se puso de manifiesto primero los enormes errores del estudio de ProPublica y, después, la complejidad del propio concepto de «fairness» (equidad) sobre el que volveré más adelante. Véase, en todo caso, primero las reflexiones críticas de FLORES, A. W.; BECHTEL, K.; LOWENKAMP, C. T.: «False Positives, False Negatives, and False Analyses: A Rejoinder to «Machine Bias: There's Software Used Across the Country to Predict Future Criminals. and it's Biased Against Blacks» en *Federal Probation,* vol. 80, núm. 2, 2016, pp. 38-46; y luego la discusión sobre los distintos elementos que incluiría el concepto de equidad que plantean KLEINBERG, J.; MULLAINATHAN, S.; RAGHAVAN, M.: «Inherent Trade-Offs in the Fair Determination of Risk Scores«. Disponible en Internet en SocArxiv, en https://arxiv.org/pdf/1609.05807.pdf (Última visita el 21/10/2018).

[70] Concretamente es la resolución 881N.W.2d749, y se puede consultar en Internet en https://www.wicourts.gov/sc/opinion/DisplayDocument.pdf?content=pdf&seqNo=171690, consultada el 10 de septiembre de 2018.

«se utilice correctamente y se observen las limitaciones y precauciones que se establecen en este documento», referidas estas esencialmente a que no sea la única herramienta en la que se apoye la decisión y que se empleen las garantías legales debidas[71].

Y COMPAS no es la única. Con propósitos similares pero un alcance geográficamente más acotado, el equivalente europeo podría ser HART (*Harm Assessment Risk Tool*) utilizada en Durham para informar sobre el riesgo de reincidencia del delincuente y, sobre la base de este pronóstico, decidir sobre la puesta en libertad o no del sujeto[72]. En ambos casos, y aunque de momento su uso es residual y requiere de una posterior supervisión humana, su generalización parece que va a más.

3.2. Riesgos asociados al uso de la IA en el sistema de justicia penal: más allá de los problemas de la «predicción», los derivados de su automatización

Cualquier disrupción tecnológica en un mundo «de tradiciones» y de evolución usualmente tan pausada como es el Derecho, suele conllevar respuestas iniciales dicotomizadas entre, por un lado el temor y la negación del cambio y, por otro, la esperanza y la aceptación del mismo sin ambages. Algo así se ha identificado ya en relación con el uso de la IA en la justicia penal: no sólo hay algo de «Hype», de exageración de las posibilidades de la IA frente a la realidad de lo que está ya aportando como se avanzó anteriormente, sino que las posiciones ante lo que esta puede suponer parecen polarizarse entre quienes, desde la esperanza de que estas herramientas ayuden a terminar con los sesgos subjetivos y con las dificultades generales para la valoración y predicción de los hechos de la justicia penal, defienden a toda costa su utilización e implementación para todo lo posible, y quienes, desde el temor a no controlar el funcionamiento de estas herramientas o a que sus efectos perniciosos para derechos fundamentales sean inevitables, prefieren renunciar a ellas[73]. Y de esa

[71] Véanse especialmente los fundamentos 8, 9, 10, 17, 28, 34, 51 y 93.

[72] Según los diseñadores, el sistema ha sido programado para extremar la cautela, y lo más probable es que clasifique a los sospechosos como de riesgo medio o alto para evitar sugerir la liberación de alguien que pueda cometer un delito. Sin embargo, también apuntan a que la herramienta es limitada en el sentido de que sólo funciona con datos de la Policía de Durham, por lo que no puede tomar en cuenta los crímenes que tuvieron lugar fuera del área. Para un análisis más detallado, véase, Oswald, M., Grace, J., Urwin, S., y Barnes, G. C.: «Algorithmic risk assessment policing models: lessons from the Durham HART model and 'Experimental' proportionality», en *Information & Communications Technology Law*, vol. 27, n.º 2, 2018, pp. 223-250.

[73] Isaac, W.: «*Hope, Hype, and Fear: The Promise and Potential Pitfalls of Artificial Intelligence in Criminal Justice*» en Ohio St. J. Crim. L., vol. 15, 2017. De modo similar Edwards distingue las actitudes de los académicos frente a la IA entre la de críticos, entusiastas y es-

polarización se puede caer en la simplificación perpetua de la supuesta bondad del término medio. Pues en realidad no se trata de aceptar simplemente que ni la IA supondrá el fin de las garantías y los derechos ni conllevará por sí misma la utópica objetivación pura de la justicia penal, sino de ir más allá y concretar con la mayor precisión qué puede aportar y qué puede poner en peligro, y derivar de ahí tanto su funcionalidad y alcance aplicativo como sus límites. Pero no ha habido tiempo para la reflexión. Como se ha visto, la utilización de la IAP se está popularizando tanto que dentro de poco es difícil pensar que el patrullaje diario no responderá a este tipo de modelos, y la IAJ está comenzando a desarrollarse pese a que no hay todavía una reflexión suficiente acerca de para qué y hasta dónde. Por eso resulta esencial poner el acento, como aquí se hará, en los riesgos y en los límites de su uso. No porque se parta de una visión temerosa, sino porque la realidad ya ha marcado que la IAJ y la IAP nos van a acompañar y debemos ser capaces de marcar cómo van a hacerlo no sólo a partir de considerar las posibilidades de su alcance sino, sobre todo, de identificar los peligros que su uso descontrolado puede conllevar.

Cuando se trata la cuestión de los riesgos del uso de la IA en el sistema de justicia penal, sin embargo, se corre el riesgo de mezclar los problemas que plantea la puesta en marcha de estas herramientas con los intrínsecamente unidos a los sistemas de valoración del riesgo. Es lógico que así sea, al fin y al cabo, y como hemos visto, la mayoría de las IAJ y algunas de las IAP no constituyen más que una evolución, en forma de automatización, de instrumentos de valoración del riesgo, y de hecho gran parte de la discusión ético-jurídica en torno a la IAJ se centra en la IAJVR y más en el aspecto actuarial de valoración del riesgo que en el proceso de automatización del mismo. Pero es importante diferenciar estos dos aspectos, y no sólo a los efectos de este trabajo que se está ocupando de la IA y, sólo en la medida que venga unido a ésta, de la valoración del riesgo de delincuencia.

En la actualidad existe una discusión académica de la máxima importancia[74] acerca de la conveniencia de reemplazar, como presupuesto central para la aplicación de cualquier medida de seguridad, la peligrosidad criminal[75], como característica psicológica individual

---

cépticos. Edwards, A.: «Big data, predictive machines and security», en McGuire, M y Holt, T. (Eds.): *The routledge handbook of technology, crime and justice,* Routledge, 2017.

[74] Véase Romeo Casabona, C.M: «Riesgo, procedimientos actuariales…», *ob. cit*. pág. 168 y ss.

[75] Urruela Mora, A.: «¿ Hacia un cambio de paradigma?: la configuración de un derecho penal de la peligrosidad mediante la progresiva introducción de medidas de seguridad aplicables a sujetos imputables en las recientes reformas penales españolas», en *Cuadernos de Política Criminal*, vol. 115, 2015, pp. 119-160; Sanz Morán, A. J.: «La peligrosidad criminal. Problemas actuales», en Carrera, E.G. (Coord.): *Delincuentes peligrosos*, 2014, pp. 61-79.

que muestra la probabilidad de que un delincuente vuelva a cometer un delito, por el riesgo de violencia (delictiva) o probabilidad objetiva de aparición de una conducta violenta determinada[76]. El debate, que a mi parecer muestra la necesidad cada vez más evidente de vincular el Derecho penal al conocimiento científico de la realidad sin que ello suponga un Derecho penal meramente empírico[77], lo es también entre visiones sobre la metodología más adecuada para la predicción de la conducta futura de una persona. Por un lado la de quienes consideran que debería haber un estudio clínico en el que primara una evaluación individualizada de la persona en concreto más allá del perfil grupal en el que la persona se encarne[78], y quienes por el contrario defienden complementar la argumentación de la peligrosidad con una fundamentación actuarial, es decir, basada en los factores de riesgo y las relaciones entre predictores y comportamiento delictivo (violento) específico demostradas empíricamente[79]. Aunque no es este el lugar para resolver esta cuestión que espero afrontar en el futuro, si quiero hacer dos consideraciones: la primera, que para resolver la cuestión del modelo de valoración de la peligrosidad/riesgo no sólo es necesario una revisión del sentido de cada una de las medidas de seguridad aplicables sino, también, la comparación de la cientificidad y la capacidad predictiva de cada una de las metodologías existentes[80]; la segunda, que el modelo de valoración del riesgo parece haber venido para quedarse y que, además, se está automatizando por medio de IA, por lo que más que negarnos a su uso y tratar de volver al sistema del juicio clínico individual, deberíamos priorizar el tratar de establecer las condiciones justas y los límites del mismo.

Porque este sí es el lugar, en cambio, para plantear que cuando las herramientas de valoración del riesgo se automatizan y se convierten en IA aún surgen más retos para la justicia penal. Aunque casi todas las herramientas actuariales tienen una parte automática (general-

---

[76] Andrés-Pueyo, A.: « Peligrosidad criminal: análisis crítico de un concepto polisémico», en Demetrio Crespo, E. (Dir.); Maroto Calatayud, M. (Coord.): *Neurociencias y Derecho Penal. Nuevas perspectivas en el ámbito de la culpabilidad y el tratamiento jurídico-penal de la peligrosidad,* Edisofer y BdeF, Madrid, 2013; Andrés-Pueyo, A.; Redondo Illescas, S.: «Predicción de la violencia: entre la peligrosidad y la valoración del riesgo de violencia», en *Papeles del psicólogo*, vol. 28, n.º 3.

[77] Miró Llinares, F.: «Hechos en tierra de normas. Una introducción epistemológica a la relevancia de la realidad fáctica en el Derecho penal», en Suárez López, J.M.; Barquín Sanz, J.; Benítez Ortúzar, I.F.; Jiménez Díaz, M. J.; Sainz-Cantero Caparrós, J. E. (Dirs.): *Estudios jurídico penales y criminológicos. En homenaje al Dr. H. C. Mult. Lorenzo Morillas Cueva,* Dykinson, Madrid, 2018.

[78] Romeo Casabona, C.M: «Riesgo, procedimientos actuariales…», *ob. cit.* pág. 167.

[79] Andrés-Pueyo, A.: « Peligrosidad criminal…, *ob. cit.*; y también Andrés-Pueyo, A.; Redondo Illescas, S.: «Predicción de la violencia: …, *ob. cit*.

[80] En este sentido sería recomendable recordar que las metodologías cuantitativas pueden ser adecuadas en términos predictivos pero pueden mejorarse, también, si a ellas se les suman otros estudios de tipo cualitativo y análisis clínicos.

mente la del cálculo), no se pueden considerar por sí mismas IA, si bien parece claro que el camino natural es su automatización y su conversión. Cuando una herramienta de valoración del riesgo pasa a ser una IA, porque se automatiza el proceso, se da un paso más que, por supuesto, tiene consecuencias más allá del supuesto incremento predictivo de la herramienta actuarial[81]. COMPAS, por ejemplo, es una IA sencilla, de tipo «*man in the loop*», pero en la que, frente a los cuestionarios de valoración del riesgo pasados «a mano» por una persona, generalmente un experto con juicio clínico, los datos no son analizados y evaluados por un ser humano, sino que son incorporados al sistema (muchos de ellos automáticamente) y evaluados de forma autónoma por la IA sin que ninguna persona haga ningún tipo de elección deliberada entre opciones. Esto plantea la problemática cuestión de si podría aceptarse que para la toma de decisiones de la justicia penal podríamos llegar a una automatización tal que finalmente fuera la máquina la que tomara la decisión, y nos obliga a reflexionar sobre el papel del ser humano. Por otro lado, las IA basadas en técnicas de Big Data requieren grandes cantidades de información para poder realizar los cálculos predictivos, lo cual plantea la cuestión de si para supuestamente incrementar la capacidad predictiva de tales herramientas vamos a aceptar poner en riesgo otros intereses dignos de tutela como la intimidad personal o la privacidad. Por último, la automatización por medio de técnicas actuales como el *machine learning,* el *deep learning* o las redes neuronales están precisamente caracterizadas por la automatización en el aprendizaje de los sistemas informáticos para el incremento de su capacidad predictiva, lo cual hace que cada vez vaya a ser más difícil que podamos explicar con claridad porqué un sistema de IA ha llegado a un cálculo, con los riesgos que ello plantea de dificultad de trazabilidad de las variables que se han tenido en consideración a la hora de la toma de decisión. Y finalmente está la cuestión de los sesgos de los datos, y de cómo se están construyendo herramientas supuestamente predictivas de la conducta de un sujeto que, sin embargo, se alimentan de información que, al menos directamente, no tienen que ver con él.

Cada uno de estos problemas de la automatización requeriría un trabajo específico. Aquí, sin embargo, sólo plantearé una visión panorámica en torno a los mismos que nos debería dar una visión de conjunto sobre las implicaciones del uso de la IA en el sistema de justicia penal.

---

[81] Aunque existe una interesante discusión sobre la mejora de las herramientas de valoración del riesgo basadas en el *machine learning* frente a las tradicionales, los estudios existentes destacan una mejora predictiva, sobre todo en procesos valorativos complejos. Véase sobre ello, con múltiples referencias, Berk, R., Heidari, H., Jabbari, S., Kearns, M., y Roth, A.: «Fairness in Criminal Justice Risk Assessments: the State of the Art», 2017. Disponible en: *arXiv preprint arXiv:1703.09207*.

3.2.1. Privacidad e IA

Las IA se alimentan de información, de grandes cantidades de datos de diferente naturaleza correspondientes con la de las variables que son tomadas en consideración en cada caso por el algoritmo predictivo. Y en comparación con tiempos anteriores en los que los datos se cedían de forma más o menos consciente, hoy la recopilación de grandes cantidades de datos referidos a múltiples esferas del actuar cotidiano de los ciudadanos es enormemente sencilla, masiva y apenas requiere de ningún tipo de comportamiento activo por parte de quien cede la información. Esto ha llevado a una preocupación creciente, cuanto menos aparente, de las administraciones públicas por la necesidad de proteger, más allá de la propia intimidad, la privacidad; esto es, al reconocimiento a nivel tanto nacional como internacional de la necesidad de la «protección de las personas físicas en relación con el tratamiento de datos personales» como un derecho fundamental[82]. De hecho, cabría no olvidar cómo, a nivel nacional, esta fundamentalidad e interés protector ya venía incorporado a nuestra Carta Magna en su artículo 18, y en particular en el 18.3 que preserva la intimidad de las comunicaciones y a el 18.4[83] que protege la intimidad informática o «habeas data»[84]. Este interés protector fue posteriormente desarrollado por la L.O. 5/1992, sustituida por la LO 15/1999, de 13 de diciembre, de Protección de Datos de Carácter personal, con una regulación más amplia y exhaustiva, que en su Exposición de Motivos daba cuenta de la necesidad de proteger la privacidad como concepto más amplio que la intimidad[85] por su vulnerabilidad ante su exposición al «desarrollo progresivo de las técnicas de recolección y almacenamiento de datos y

---

[82] Así lo dice la Directiva (EU) 2016/679 en su primer considerando; el artículo 8, apartado 1, de la Carta de los Derechos Fundamentales de la Unión Europea («la Carta») y el artículo 16, apartado 1, del Tratado de Funcionamiento de la Unión Europea (TFUE) establecen que toda persona tiene derecho a la protección de los datos de carácter personal que le conciernan.

[83] Artículo 18.3: «Se garantiza el secreto de las comunicaciones y, en especial, de las postales, telegráficas y telefónicas, salvo resolución judicial». Artículo 18.4: «La ley limitará el uso de la informática para garantizar el honor y la intimidad personal y familiar de los ciudadanos y el pleno ejercicio de sus derechos».

[84] López Barja de Quiroja, J.: «La intimidad después del a Reforma del artículo 197 del Código Penal: La divulgación sin consentimiento de imágenes o grabaciones obtenidas con consentimiento», en Bacigalupo Saggese, S., Feijoo Sánchez, B., y Echano Basaldua, J. I. (Coord.), *Estudios de Derecho Penal. Homenaje al profesor Miguel Bajo*, Editorial Universitaria Ramón Areces, Madrid, 2016, pp. 1024 y ss.

[85] De hecho, en la propia Exposición de Motivos el legislador conceptualiza la privacidad como «un conjunto, más amplio, más global, de facetas de su personalidad que, aisladamente consideradas, pueden carecer de significación intrínseca pero que, coherentemente enlazadas entre sí, arrojan como precipitado un retrato de la personalidad del individuo que éste tiene derecho a mantener reservado. Y si la intimidad, en sentido estricto, está suficientemente protegida por las previsiones de los tres primeros párrafos del artículo 18 de la Constitución y por las leyes que los desarrollan, la privacidad puede

de acceso a los mismos» y que, tal y como afirma Gutiérrez Francés, «abrió una nueva etapa con el reconocimiento y tutela de la privacidad, primero en el orden administrativo, y luego, a partir de 1995, refrendado en el orden penal»[86]. A pesar de que la intimidad, *stricto sensu*, y la privacidad en un sentido más amplio, ha conllevado largas discusiones e interpretaciones dogmáticas[87], hoy en día resulta innegable para la doctrina que la intimidad, como unidad mínima por su conexión con la dignidad y el libre desarrollo de la personalidad[88], constituye un derecho fundamental del individuo. Y esto porque la misma concede ciertas esferas al individuo indispensables para el correcto desenvolvimiento en sociedad al configurarse como aquella «parcela de la vida que se considera no sólo secreta sino absolutamente privada y reservada para uno mismo»[89] y que lleva intrínsecamente consigo la posibilidad de excluir a otros de dicha parcela. Y lo mismo cabría decir del derecho a la privacidad y, por extensión, de privacidad informática, al plantear un concepto que incluye la posibilidad de disponer de aquellos datos que, no siendo propios del núcleo duro de la intimidad, como podrían ser

---

resultar menoscabada por la utilización de las tecnologías informáticas de tan reciente desarrollo».

[86] Gutiérrez Francés, M. L.: «La privacidad en el espacio virtual (riesgos y cauces de protección)», en Velásquez Velásquez, F., Vargas Lozano, R., y Jaramillo Restrepo, J. D., *Derecho penal y nuevas tecnologías. A propósito del Título VII Bis del Código Penal*, Universidad Sergio Arboleda, Bogotá, D. C., 2016, p. 125. Así, tal y como explica Morón Lerma, E.: «Intención del agresor y ataque a la intimidad», en Quintero Olivares, G., y Morales Prats, F. (Coord.), *El Nuevo Derecho Penal Español. Estudios Penales en Memoria del Profesor José Manuel Valle Muñiz*, Aranzadi, Elcano, Navarra, 2001, pp. 1609 y ss., la doctrina apreció positivamente la sistemática de la criminalización de los delitos contra la intimidad en el Código penal en comparación a la anterior regulación, dando respuesta a casos que anteriormente habían quedado sin tutela a pesar de que vulneraban flagrantemente este derecho personalísimo al que la Constitución Española le había dado carta de naturaleza de derecho fundamental. Por ejemplo, uno de estos supuestos sería, tal y como lo reseña Gutiérrez Frances, M. L.: «La privacidad…», *ob. cit.*, el de «PubliGest».

[87] Véase profusamente Rueda Martín, M. A.: *La nueva protección de la vida privada y de los sistemas de información en el Código Penal*, Atelier, Barcelona, 2018.

[88] Así, la STC núm. 17/2013, de 31 de enero recuerda que «constituye doctrina consolidada de este Tribunal que el derecho a la intimidad personal garantizado por el art. 18 CE, estrechamente vinculado con el respeto a la dignidad de la persona (art. 10.1 CE), implica la existencia de un ámbito propio y reservado frente a la acción y el conocimiento de los demás, necesario, según las pautas de nuestra cultura, para mantener una calidad mínima de la vida humana. Además el art. 18 CE confiere a la persona el poder jurídico de imponer a terceros, sean estos poderes públicos o simples particulares (STC 85/2003), de 8 de mayo, F. 21), el deber de abstenerse de toda intromisión en la esfera íntima y la prohibición de hacer uso de lo así conocido, y de ello se deduce que el derecho fundamental a la intimidad personal otorga cuanto menos una facultad negativa o de exclusión, que impone a terceros el deber de abstención de intromisiones salvo que estén fundadas en una previsión legal que tenga justificación constitucional y que sea proporcionada, o que exista un consentimiento eficaz que lo autorice, pues corresponde a cada persona acotar el ámbito de intimidad personal y familiar que reserva al conocimiento ajeno (STC 206/2007, de 24 de septiembre, F. 5, por todas)».

[89] López Barja de Quiroga, J.: «La intimidad…», *ob. cit.*, p. 1023.

aquellos que desvelan las características más personales del individuo, lo hacen respecto de otras características que, unidas todas ellas de manera coherente, podrían dar lugar a la concreción y perfilación de alguien específico[90], sobre lo que volveré más adelante dado que sobre ello versará el siguiente punto de este trabajo.

Sin embargo, es ese mismo incremento exponencial de las posibilidades que hoy ofrecen las tecnologías para la recogida y análisis de grandes cantidades de datos y su uso por parte de las administraciones lo que lleva a escenarios de posibles afectaciones, e incluso afectaciones «supuestamente legítimas» por parte del Estado, que subvierten el inicial interés limitador que los derechos fundamentales tienen frente a los poderes públicos, y que ahora han acabado socavando los argumentos de seguridad, prevención e investigación del delito[91]. Dicho de otro modo, no es algo nuevo ver cómo la seguridad del Estado, la defensa o la seguridad pública han terminado por viciar progresivamente aquella voluntad germinal de proteger la privacidad y la intimidad como aquel derecho fundamental que abre y atraviesa la Directiva 2016/679 con el desarrollo del Reglamento General de Protección de Datos, y que ha dado como resultado numerosas referencias en las que se explicitan excepcionalidades para los que el tratamiento total o parcialmente automatizado de datos personales no tiene por qué estar sujeto estrictamente al reglamento[92].

Parece, pues, que el tratamiento automatizado, semiautomatizado o manual de datos personales por los sistemas policiales o judiciales europeos vaya a estar más sujeto a una lógica de carta blanca y de afectación a la privacidad indiscriminada bajo el pretexto de una lucha eficaz contra la delincuencia, que a una protección efectiva de los ciudadanos.

---

[90] De hecho, tal y como explica López Barja Quiroga, J., «La intimidad…», *ob. cit.*, p. 1024, el derecho a la privacidad que surge como concepto anglosajón y elaborado por el juez americano Cooley, surge como el «derecho a ser dejado en paz», y que sugiere la posibilidad de controlar por el individuo todo aquello que concierne a todos los aspectos de su intimidad.

[91] Tal y como afirma Gutiérrez Francés, M. L., «La privacidad...», *ob. cit.*, p. 127, el volumen de información que existe sobre el ciudadano en el ciberespacio es cada vez mayor; el contenido de dicha información es cada vez más completo, «comprendiendo ahora datos más sensibles y personalísimos», y, dicha información es vulnerable ante conductas ilegítimas no solo respecto de otros individuales o empresas, sino del propio Estado quien cada vez dispone de más información sobre la ciudadanía.

[92] De acuerdo con el Art. 2.2 de la Directiva (EU) 2016/679: El presente Reglamento no se aplica al tratamiento de datos personales: a) en el ejercicio de una actividad no comprendida en el ámbito de aplicación del Derecho de la Unión; b) por parte de los Estados miembros cuando lleven a cabo actividades comprendidas en el ámbito de aplicación del capítulo 2 del título V del TUE; c) efectuado por una persona física en el ejercicio de actividades exclusivamente personales o domésticas; d) *por parte de las autoridades competentes con fines de prevención, investigación, detección o enjuiciamiento de infracciones penales, o de ejecución de sanciones penales, incluida la de protección frente a amenazas a la seguridad pública y su prevención.* Las cursivas son nuestras.

Directivas europeas como la 2016/680, del Parlamento Europeo y del Consejo, de 27 de abril de 2016, relativa a la protección de las personas físicas en lo que respecta al tratamiento de datos personales por parte de las autoridades competentes para fines de prevención, investigación, detección o enjuiciamiento de infracciones penales o de ejecución de sanciones penales, y a la libre circulación de dichos datos y por la que se deroga la Decisión Marco 2008/977/JAI del Consejo, se han acabado convirtiendo en instrumentos jurídicos vinculantes a todos los Estados miembros que, más allá de ensayar nuevas fórmulas para mejorar y facilitar el trabajo de las fuerzas policiales y judiciales en el intercambio de información, han acabado legitimando un tratamiento de datos personales cada vez más diversos, aparentemente indirectos, pero que formando parte de un conjunto de datos mayor y más estructurado incrementa significativamente la capacidad de las agencias de seguridad de identificar a individuos concretos. Y lo mismo cabrá esperar del futuro Reglamento *e-Privacy*[93] que, pese a su aparente rigidez[94], ya establece en su propuesta que serán posibles las excepciones conforme a fines legítimos.

En nuestro país, encontramos la excepcionalidad a la protección de la privacidad de los ciudadanos en la LO 13/2015, de 5 de octubre, de modificación de la Ley de Enjuiciamiento Criminal para el fortalecimiento de las garantías procesales y regulación de las medidas de investigación tecnológica. Bajo el pretexto de que «renovadas formas de delincuencia ligadas al uso de las nuevas tecnologías han puesto de manifiesto la insuficiencia de un cuadro normativo concebido para tiempos bien distintos», el legislador habilita a nuestras agencias de seguridad para el ejercicio de la intromisión en datos personales con el objetivo de descubrir e investigar la comisión de delitos, proporcionando poderosas herramientas de investigación a los poderes públicos. En este sentido, si bien las nuevas previsiones que configuran lo que, con acierto, Cuerda Arnau ha denominado la «tecnovigilancia»[95] han sido en parte bien recibidas por la doctrina[96] puede que por su mera

[93] Véase la Propuesta de Reglamento del Parlamento Europeo y del Consejo, sobre el respeto a la vida privada y la protección de los datos personales en el sector de las comunicaciones electrónicas y por el que se deroga la Directiva 2002/58/CE (Reglamento sobre la privacidad y las comunicaciones electrónicas). Disponible en Internet en: https://eur-lex.europa.eu/legal-content/ES/ALL/?uri=CELEX%3A52017PC0010.

[94] Ortiz López, P., «Regulando la privacidad del futuro. Análisis de la Propuesta de Reglamento Europeo de e-Privacy y su interconexión con el Reglamento General de Protección de Datos», en *Diario La Ley*, n.º 10, 2017.

[95] Cuerda Arnau, M. L.: «La reforma de la Ley de Enjuiciamiento Criminal en materia de tecnovigilancia. Visión de conjunto», en Alonso Rimo, A., Cuerda Arnau, M. L., y Fernández Hernández, A. (Dirs.), *Terrorismo, sistema penal y derechos fundamentales*, Tirant lo Blanch, Valencia, 2018, pp. 509 y ss.

[96] Ibidem.; Ridaura Martínez, M. J., «El legislador ausente del artículo 18.3 de la Constitución (la construcción pretoriana del derecho al secreto de las comunicaciones)», en *Revista de Derecho Político*, n.º 100, 2017, p. 391.

positivización[97], no han pasado inadvertidas las previsiones que fácilmente pueden sucumbir a la ausencia de las garantías propias de un Estado Social y Democrático de Derecho en el acceso y tratamiento de los datos personales. Me refiero a como bien advierte CUERDA ARNAU los peligros inherentes «a las vigilancias sistemáticas que, al amparo de labores de investigación, terminan por ser un modo de recopilación de datos de ciudadanos que se limitan a ejercer derechos fundamentales»[98]. Y todo ello en atención a tres extremos habilitantes que establece la Ley: 1) el establecimiento como presupuesto suficiente para legitimar las intromisiones en la intimidad y la privacidad por parte de los poderes del Estado basado en la necesidad de investigación ya no de posibles delitos de terrorismo, sino de cualquier delito cometido a través de instrumentos informáticos o cualquier otra tecnología de la información o la comunicación, de conformidad con el artículo 588 ter a., para el caso de las interceptaciones de las comunicaciones telefónicas y telemáticas o el artículo 588 septiés a., para los registros remotos sobre equipos informáticos; 2) los plazos excesivos previstos para este tipo de intromisiones que pueden llegar hasta los dos años[99], y 3) por las excepciones bajo la denominación de «urgente», que permite la intromisión en las comunicaciones sin autorización judicial previa tensionando así la regla general que más garantías ofrece al ciudadano[100].

Pero, ¿qué tipo de datos personales pueden ser recopilados y utilizados con estos fines? Ciertamente, y reitero que con un carácter vinculante a todos los Estados miembros, la Directiva 2016/679 no deja ninguna duda acerca de lo que debemos entender tanto como datos personales a proteger, como a vulnerar en situaciones de excepcionalidad: a saber,

---

[97] Ya que anteriormente no estaba prevista la intromisión en las comunicaciones telemáticas. Dicha ausencia permitía la comisión de determinadas arbitrariedades utilizada por los poderes de investigación para intervenir en todo tipo de comunicaciones sin haber previsión legal que así lo estableciera. Dicho de otro modo, tal y como explica LÓPEZ-BARAJAS PERA, I.: «Garantías constitucionales en la investigación tecnológica del delito: previsión legal y calidad de la ley», en *Revista de Derecho Político*, n.º 98, 2017, p. 114, «la comunicación telemática deja de tratarse como accesoria o instrumental de la comunicación telefónica».

[98] CUERDA ARNAU, M. L.: «La reforma…», *ob. cit.*, p. 523.

[99] El art. 588 ter g. LECrim establece que la intervención de las comunicaciones tendrá un plazo inicial de tres meses, pero se podrá prorrogar hasta dos años. Lo mismo sucede con la medida de utilización de dispositivos de captación de la imagen, de seguimiento y de localización tal y como advierte BUENO DE MATA, F.: «Comentarios y reflexiones sobre la Ley Orgánica 13/2015 de modificación de la Ley de Enjuiciamiento Criminal para el fortalecimiento de las garantías procesales y la regulación de las medidas de investigación tecnológica», en *Diario La Ley*, n.º 8792, 2016.

[100] GONZÁLEZ NAVARRO, A.: «El uso de nuevas tecnologías en la investigación de delitos de terrorismo», en ALONSO RIMO, A., CUERDA ARNAU, M. L., y FERNÁNDEZ HERNÁNDEZ, A. (DIRS.), *Terrorismo, sistema penal y derechos fundamentales*, Tirant lo Blanch, Valencia, 2018, p. 558.

«toda información sobre una persona física identificada o identificable; se considerará persona física identificable toda persona cuya identidad pueda determinarse, directa o indirectamente, en particular mediante un identificador, como por ejemplo un nombre, un número de identificación, datos de localización, un identificador en línea o uno o varios elementos propios de la identidad física, fisiológica, genética, psíquica, económica, cultural o social de dicha persona». Y pese a que esto dota al Reglamento de una mayor seguridad jurídica, a mi parecer, la tendencia se invierte cuando hablamos de los metadatos aplicados a la identificación de sujetos por medio de su huella digital, inspiradoras en gran parte de las arquitecturas de las IA actuales para la investigación criminal *online,* y que la directiva pretende saturar con la expresión de «identificadores en línea». Bajo esta denominación se pueden incluir numerosas fuentes de nuevos tipos datos derivados de la actividad de los usuarios con las TIC, y que integran en sus procesos analíticos variables como por ejemplo el medio de creación de los datos, finalidad de los datos, hora y fecha de creación, creador o autor de los datos, ubicación en una red informática donde se crearon los datos, normas utilizadas, tamaño de archivo, calidad de los datos, fuente de los datos, proceso utilizado para crear los datos, entre otros metadatos, que han mostrado una igual o superior capacidad para la identificación de sujetos concretos[101]. Esta miopía técnica del legislador podría quedar parcialmente justificada ante las enormes dificultades existentes para la cuantificación y calificación de un número de atributos o metadatos altamente fluctuante y que es resultado de las constantes actualizaciones de las plataformas digitales, especialmente de redes sociales[102], y cuya inobservancia, sea cual sea el fin al que respondan, puede comprometer la correcta aplicabilidad de los principios básicos de protección de datos, esto es, la legalidad, equidad y transparencia, tan relevantes en un ámbito tan sensible como es el de la prevención e investigación del delito.

Es por ello que, ante este escenario caracterizado por un incremento significativo de los datos, los cuales son más y más diversos hoy, se invoca la necesidad de recuperar la máxima teleológica de que la privacidad y la intimidad de las personas merece una especial protección gubernamental. Y esto es algo que en la actualidad ha llevado al desarrollo de algunas herramientas o procedimientos de protección de la privacidad de enorme interés. Entre ellos, como bien analiza Valls[103], uno de los prin-

---

[101] Sekara, V., Mones, E., y Jonsson, H: *Temporal Limits of Privacy in Human Behavior*, 2018. Disponible en: *arXiv preprint arXiv:1806.03615*.

[102] Ghosh, I., y Singh, V. K.: «Predicting privacy attitudes using phone metadata», en Xu, K., Reitter, D., Lee, D., Osgood, N. (Eds), *Social, Cultural, and Behavioural Modeling*, SBP-BRiMS 2016. Lecture Notes in Computer Science, vol. 9708, Springer, Cham, 2016, pp.51-60.

[103] Sobre el alcance y fundamento jurídico del *Privacy Impact Assessment*, cfr. Valls Prieto, J.: *Problemas jurídico penales…*, *ob. cit.*, y Valls Prieto, J.: «El uso de inteligencia artificial…», *ob. cit.*, pp. 77-106.

cipales instrumentos para la evaluación del impacto en la protección de la privacidad es el PIA (*Privacy Impact Assessment*)[104], el cual refiere a la obligación del responsable del tratamiento de datos de efectuar una evaluación de impacto y de documentarla antes de iniciar el tratamiento de datos previsto, muy especialmente en tres supuestos estrechamente relacionados con el uso de IA y que pueden resultar en un alto riesgo para los derechos y libertades de las personas. Hablamos más concretamente de 1) una evaluación sistemática y extensa de los aspectos personales de un individuo, incluido el perfil; 2) procesamiento de datos confidenciales a gran escala y el 3) monitoreo sistemático de áreas públicas[105].

3.2.2. IA y sistema penal «justo»: especial atención a la discriminación algorítmica

La introducción y utilización de las IA en el sistema de justicia penal ha llevado a que algunas voces adviertan de otros riesgos relacionados con su uso, distintos a la potencial afectación a la privacidad, que podrían surgir y perturbar la propia eficacia del sistema de justicia y afectar a derechos fundamentales de las personas[106]. Suscitan especial preocupación las garantías procesales que podrían verse afectadas por la utilización durante el proceso penal de herramientas de este tipo. Así sucede, por ejemplo, con el Derecho de defensa, que podría entrar en colisión con la falta de publicidad de todos o algunos elementos del algoritmo impidiendo, así, una adecuada defensa frente a las atribuciones o valoraciones realizadas[107]. Pese a que en EE.UU. esta cuestión fue planteada en relación con la IAJVR COMPAS en el caso State v Loomis y el Tribunal Supremo de Wisconsin decidió que no había violación del derecho de defensa pese a impedir el acceso al algoritmo por consideraciones de propiedad intelecutal[108], creo que acierta la doctrina española que ha analizado la cuestión al considerar que conforme a nuestro Ordenamiento Jurídico no sería aceptable la utilización de herramientas en la justicia que los litigantes no estuvieran en disposición de conocer[109].

---

[104] Véanse Artículos 35 y 36 y considerandos (89) a (96) del GDPR.

[105] Véase, con gran profundidad, el desarrollo y explicación que realiza VALLS, PRIETO, J.: *Problemas jurídico penales…*, *ob. cit.*, pp. 85 y ss.

[106] Plantea NIEVA FENOLL la cuestión prestando atención a derechos como el derecho al juez imparcial, de defensa, a la intimidad y a la presunción de inocencia. *Inteligencia artificial y proceso…,* NIEVA FENOLL, J.: Inteligencia artificial y proceso …, *ob. cit.*, pp. 127 y ss.

[107] NIEVA FENOLL, J.: Inteligencia artificial y proceso …, *ob. cit.*, p. 141.

[108] Véase el análisis sobre esta resolución de NIEVA FENOLL, J.: Inteligencia artificial y proceso …, *ob. cit.*, pp. 70 y ss., y 140 y ss.

[109] En este sentido recientemente, tanto ROMEO CASABONA, C.M: «Riesgo, procedimientos actuariales…», *ob. cit.*, p. 178 y s., como NIEVA FENOLL, J.: Inteligencia artificial y proceso …, *ob. cit.*, pp. 141 y ss.

Sin embargo, si hay una cuestión que ha suscitado especial preocupación y que ha ocupado a la academia en los últimos tres años es la potencial afectación al funcionamiento equitativo, en el sentido de no discriminatorio, del sistema de justicia en un Estado Social y Democrático de Derecho. La aplicación equitativa de la justicia penal podría verse en riesgo por la potencial discriminación algorítmica derivada de la utilización de la IA para la perfilación de individuos, de colectivos, de áreas urbanas; para la búsqueda de mensajes de radicalización o la identificación de mensajes radicales; para la distribución de los recursos policiales y del patrullaje; para la selección de los lugares donde se ubican los radares de velocidad, etc. El riesgo aquí ya no es sólo que se acceda a datos sensibles de carácter personal para poder configurar adecuadamente tales algoritmos, sino que al hacerlo sobre datos específicos y, por tanto, sesgados, las herramientas actuariales también respondan a los mismos o creen nuevos sesgos debido a una mala interpretación de la realidad. De nuevo estos riesgos ya se pusieron de manifiesto en relación con el sistema COMPAS y el caso «State v. Loomis», en el que no se atendió a la alegación del apelante relativa a que en su valoración la IA había dado mucha importancia al género y se había actuado de forma sesgada. Dejando a un lado esta resolución, que no dio la razón al apelante a partir del doble argumento de que el uso del género por parte de la IA promueve la precisión de la herramienta y de que el apelante no cumplió con su carga de demostrar que el tribunal de sentencia realmente se basó en este factor y no en otros en la sentencia[110], lo cierto es que la cuestión no es sencilla. Como se verá, una importante parte de los sesgos algorítmicos puede devenir de los propios patrones no aleatorios en los que se plasma la realidad en cuestiones como la edad, el género, o la etnia; mientras que otros pueden producirse por defectos analíticos por la falta de datos que, sin embargo, podrían no ser peores que los que se producen cuando el análisis para la toma de decisión lo llevan a cabo personas sin el uso de sistemas informáticos.

Por ello, y como ya he señalado en otro lugar[111], hay que comenzar por reconocer que estamos ante uno de los debates éticos más complejos e importantes a los que se van a enfrentar los sistemas de justicia penal en las próximas décadas[112] conforme, y parece que de manera imparable, se vayan automatizando cada vez más las decisiones en materia de prevención e investigación del delito. Sin ánimo simplificador pero sí explicativo, podemos comenzar por señalar que habría dos posiciones enfrentadas iniciales. Por un lado, la de quienes defienden que los algo-

[110] State v. Loomis, considerandos 75 a 86.

[111] Miró Llinares, F.: «Apuntes sobre la relación entre Derecho penal e Inteligencia Artificial», en *Libro Homenaje al Profesor Gonzalo Quintero Olivares*. En prensa.

[112] Para una introducción a la problemática de la discriminación algorítmica en EE.UU., véase Flores, A. W.; Bechtel, K.; Lowenkamp, C. T.: «False Positives,…» , *ob. cit.*, pp. 38 y ss.

ritmos no deben renunciar a determinadas variables como la edad, la etnia, el grupo social y otras demográficas con la excusa de la afectación a los derechos de otros, ya que la exclusión de variables altamente correlacionadas, pero a primera vista éticamente cuestionables, afectaría la capacidad predictiva de la herramienta y con ello aumentaría el riesgo de su uso viciado[113]. Por otro lado nos encontraríamos con quienes defienden que las cargas legales deben siempre tener alguna relación con la responsabilidad individual más que con otros factores que son problemáticos desde una perspectiva ética o legal y que no dependen del individuo mismo, tal y como sucede con las variables sociodemográficas propias de las herramientas actuariales[114], por lo que las mismas nunca debieran ser utilizadas en este tipo de algoritmos usados por la IA para distintos fines relacionados con la justicia penal. En algún lugar entre ellas, o en lo que ambas visiones pueden resultar acertadas, debiéramos encontrar el camino por el que debería transitar un uso ético de los algoritmos informáticos en la IA. Realizaré a continuación algunas reflexiones que, a mi parecer, deben ser tomadas en consideración en relación con el debate respecto al uso de los sistemas de inteligencia artificial en la justicia penal.

La primera reflexión consiste en la necesidad de diferenciar dos tipos principales de sesgos en los algoritmos propios de las IA: sesgos en los datos de entrenamiento y sesgos por una distribución desigual real de las variables[115]. Respecto al primer grupo, las IA actuales se basan en datos de entrenamiento dentro de modelos de aprendizaje automático. Dicho de otro modo, en contraste con aquellos algoritmos basados en reglas precodificadas y completamente especificadas, las IA aprenden con nuevos ejemplos[116]. Es por esto por lo que se ha hablado de un

---

[113] Cfr. Slobogin, C.: *Proving the unprovable: The role of law, science, and speculation in adjudicating culpability and dangerousness*, Oxford University Press, 2006; Larkin, P.: «Managing prisons by the numbers: Using the good-time laws and risk-needs assessments to manage the federal prison population», en *Harvard Jounal of Law and Public Policy,* vol. 1, n.º 1, 2014, pp. 1-29.

[114] Cfr. Berk, R., y Bleich, J.: «Forecasts of violence to inform sentencing decisions», en *Journal of Quantitative Criminology*, vol. 30, n.º 1, 2014, pp. 79-96; Netter, B.: «Using Groups Statistics to Sentence Individual Criminals: An Ethical and Statistical Critique of the Virginia Risk Assessment Program», en *J. Crim. L. & Criminology*, vol. 97, 2006, pp. 699 y ss.; Zinger, I.: Zinger, I.: «Actuarial Risk Assessment and Human Rights: A Commentary», en *Canadian Journal of Criminology and Criminal Justice*, vol. 46, n.º 5, 2004, pp. 607-620.

[115] Para una contextualización de este tipo de sesgos en el marco de la política europea de antidiscriminación, véase el reciente trabajo de Hacker, P.: «Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies Against Algorithmic Discrimination Under EU Law», en *Common Market Law Review*, 2018. Disponible en Internet en: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3164973.

[116] Véase una aproximación divulgativa a la idea de autoaprendizaje de algoritmos en Jones, N.: «Computer science: The learning machines», en *Nature News*, vol. 505, n.º 7482, 2014, pp. 146 y ss.

triple riesgo de sesgo con los datos de entrenamiento que se perpetúan con el autoaprendizaje del algoritmo: 1) que la máquina haya sido «alimentada» con datos erróneamente calificados por el programador, lo que viciaría el aprendizaje desde el comienzo[117]; 2) que el set de datos inicial sea resultado de un muestro no representativo, reduciendo su validez y aumentado el error de la interpretación de los resultados[118] y 3) que los parámetros del aprendizaje se realicen sobre intervalos temporales fuertemente acotados[119]. Lo cierto es que la inobservancia de este tipo de limitaciones acerca de los datos utilizados en estos algoritmos predictivos perpetúa, en efecto, la discriminación algorítmica hacia los grupos infrarrepresentados, pudiendo lesionar o poner en riesgo sus derechos fundamentales[120]. Un problema que acaba acentuándose cuando el tratamiento de grandes cantidades deja al margen otras cuestiones fundamentales en los procesos de toma decisiones basadas en evidencias relativas a los procesos de medición, validez, dependencias entre los elementos de nuestros datos, marco teórico u objetivos a los que pretende responder el procesamiento de datos, entre otros aspectos epistemológicos claves[121]. Resulta esencial, pues, eliminar estos sesgos de las IAJ e IAP o, al menos, conocer las limitaciones de los algoritmos derivadas de ellos, lo que obliga, a mi parecer, a exigir que los mismos

---

[117] Aplicado específicamente a las IA, Barocas, S., y Selbst, A. D.: «Big data's disparate impact», en *Cal. L. Rev.*,vol. 104, 2016, pp. 671 y ss. Para una perspectiva general sobre sesgos, además véase Greenwald, A. G., y Krieger, L. H.: «Implicit bias: Scientific foundations», en *California Law Review*, vol. 94, n.º 4, 2006, pp. 945-967.

[118] Véase Calders, T., y Žliobait, I.: «Why unbiased computational processes can lead to discriminative decision procedure», en Custers, B., *et al.* (Eds.), *Discrimination and privacy in the information society*, Springer, Berlin, 2013, pp. 43-57.

[119] Como ejemplo de este tipo de sesgos, Caliskan, A., Bryson, J. J., y Narayanan, A: «Semantics derived automatically from language corpora contain human-like biases», en *Science*, vol. 356, n.º 6334, 2017, pp. 183-186.

[120] Entre ellas, si los datos de entrenamiento reflejan necesariamente prácticas pasadas, una de las estrategias más recomendadas es que los algorítmicos se actualicen regularmente utilizando los datos más recientes; para una propuesta véase Silberman, Gabriel M., et al: «Detecting and reducing bias (including discrimination) in an automated decision-making process», U.S. Patent Application No. 15/595,220, 2017. Pero incluso estas medidas de prevención de riesgos éticos incluyen nuevas preguntas no resueltas: ¿Hay un rol para el aprendizaje en línea?, ¿cuántos datos históricos deben descartarse a medida que se revisan los datos de entrenamiento?, ¿se debería dar más peso a los datos de entrenamiento más recientes en el análisis?, etc. Por su parte, Shneidermana, en una línea similar pero con un alcance económico adicional, plantea un propuesta de supervisión de IA aplicadas al sistema de justicia penal desde una triple perspectiva: 1) un modelo de junta de revisión en la que los proveedores o las agencias deben presentar su herramienta o algoritmo antes de cualquier implementación en el mundo real; 2) una monitorización continua que recuerde a las compañías y fundaciones sin fines de lucro lo que deben; y 3) un análisis retrospectivo de escenarios de «desastre» (Shneidermana, B.: «The dangers of faulty, biased, or malicious algorithms requires independent oversight», en *PNAS*, vol. 113, n.º 48, 2016).

[121] Lazer, D., Kennedy, R., King, G., y Vespignani, A.: «The parable of Google Flu: traps in big data analysis», en *Science*, vol. 343, n.º 6176, 2014, pp. 1203-1205.

sean públicos y a que puedan realizarse evaluaciones técnicas que permitan, incluso durante el proceso y como parte del ejercicio del derecho de defensa, poner en duda los resultados.

Pero incluso evitando el primer conjunto de sesgos debemos considerar un segundo grupo que viene marcado por la existencia de variables que verdaderamente se distribuyen de manera desigual en la realidad. Me refiero más concretamente a factores aparentemente neutrales recogidos por la IA[122] pero que no se refieren al actuar de la persona en concreto sino al actuar pasado de sus iguales, aquellos que comparten género, etnia, edad, etc. Se trata de caracteres «inmutables» o casi inmutables, de los que el sujeto no se puede desprender[123]. Al respecto se discute si las IA, especialmente las de valoración del riesgo, pueden trabajar con ese tipo de variables o por el contrario deberían ser retiradas de los algoritmos. Como se ha adelantado, la cuestión no es sencilla. Pensemos en los factores relacionados con el género. Resulta difícil pensar en una IAVR que no tenga en cuenta el género de la persona a la hora de hacer la evaluación predictiva, teniendo en cuenta que las tasas de reincidencia pueden ser muy diferentes (dependiendo del delito) según el mismo. Y los estudios empíricos demuestran que la diferencias de base en todos estos «factores protegidos» son altísimas, «como ocurre por ejemplo con el hecho de ser hombre joven y su relación con la causación de delitos violentos»[124]. Por eso señala HAMILTON que «cualquier herramienta de evaluación de riesgos que no distinga entre hombres y mujeres clasificará erróneamente a ambos sexos»[125], y de ahí que la sentencia del caso State v. Loomis, señalara que «si la inclusión del género promueve la precisión del sistema de IA, entonces está sirviendo a los intereses de las instituciones y los acusados, en lugar de a un propósito discriminatorio»[126].

El argumento que hay detrás de mantener los indicadores referidos a este tipo de datos estriba, pues, en la exigencia de *accurancy* (exactitud) o *validity* (validez), uno de los elementos esenciales que se debe exigir a un algoritmo predictivo[127]. Pero todos los autores que se han dedicado a

---

[122] Cfr. ROMEI, A., y RUGGIERI, S.: «A multidisciplinary survey on discrimination analysis», en *The Knowledge Engineering Review*, vol. 29, n.º 5, 2014, pp. 582-638; HACKER, P., y PETKOVA, B.: «Reining in the big promise of big data: transparency, inequality, and new regulatory frontiers», en *Nw. J. Tech. & Intell. Prop.*, vol. 15, 2017.

[123] SLOBOGIN, C.: *Proving the unprovable:…, ob. cit.*

[124] BERK, R., HEIDARI, H., JABBARI, S., KEARNS, M., y ROTH, A.: «Fairness in Criminal Justice Risk Assessments:…», *ob. cit.*

[125] HAMILTON, M.: «Risk-Needs Assessment: Constitutional and Ethical Challenges», n.º 52, *Am. Crim. L. Rev.,* 2015.

[126] State v. Loomis, considerando 82.

[127] Véase sobre ello SLOBOGIN, C. «Principles of Risk Assessment: Sentencing and Policing», en *Ohio St. J. Crim. L.*, vol. 15, 2017. Junto a la validez incluye SLOBOGIN, el principio de ajuste y el de equidad.

analizar estas herramientas están de acuerdo en que un instrumento valorativo de este tipo no sólo debiera ser científicamente válido o exacto, sino equitativo, y hay autores que han puesto en duda que lo sean aquellas que utilicen este tipo de variables inmutables.

Así lo defienden autores como NETTER[128] o STARR[129], quien define el uso de las herramientas que incluyan factores democráticos como «la aceptación explícita de la condena discriminatoria saneada por un lenguaje científico». La autora señala que la utilización de tales variables supone una violación del principio de igualdad, particularmente de la exigencia derivada de él de que las personas tienen derecho a ser tratadas como individuos y a ser responsabilizadas por sus hechos y no por los de los demás[130]. De fondo lo que hay, y por eso creo que la resolución de esta cuestión requiere una reflexión mayor, es la duda de si la utilización de estas características no conllevaría una suerte de perversión del concepto de «culpabilidad», clave en nuestra construcción de la Teoría del Delito[131], ya que haría posible establecer la responsabilidad penal a partir de la evaluación de características de los sujetos afectados sobre las que ellos no son responsables[132] o, en cierto sentido, tienen un control muy bajo, como en los casos de una enfermedad física o mental[133]. Y a estos argumentos ético-jurídicos podríamos unir otros relacionados con la perpetuación de los sesgos, lo que han denominado «*stakeholders sensitivities*» o «sensibilidades de los grupos de interés»[134], o los costes sociales derivados de la discriminación algorítmica, pues se ha señalado que estos deberían ser inasumibles en

---

[128] NETTER, B.: «Using Groups Statistics …», *ob. cit.*, p.728.

[129] STARR, S.: «Evidence-Based Sentencing and the Scientific Rationalization of Discrimination», en *Stan. L. Rev.,* 2014. Añade la autora otras razones para rechazar estos instrumentos. En concreto, y frente al argumento de que la utilización de tales factores puede estar justificada para asegurar la fiabilidad de los instrumentos, pone en duda tanto la precisión de los instrumentos, afirmando que los mismos no superan las predicciones informales de los jueces, como la supuesta superación de los sesgos discriminatorios que se atribuyen a los tribunales.

[130] STARR, S.: «Evidence-Based Sentencing…», *ob. cit.*, pág. 828 y ss. En sentido similar señala ZINGER, I.: «Actuarial Risk Assessment…», *ob. cit.*, p. 611, que los principales marcos jurídicos internacionales muestran una tendencia hacia un trato equitativo de las personas y prohíben explícitamente cualquier práctica discriminatoria, lo que podría incluir, entre otros elementos, el uso de estas características. ZINGER, I.: «Actuarial Risk Assessment…», *ob. cit.*, pp. 842 y ss.

[131] Véase sobre todo ello las reflexiones recientes, en un sentido muy crítico, de ROMEO CASABONA, C.M: «Riesgo, procedimientos actuariales…», *ob. cit.* pág. 166 y ss, especialmente 178 y ss.

[132] TONRY, M.: «Legal and Ethical Issues in the Prediction of Recidivism», en *Federal Sentencing Reporter,* vol. 26, n.º 3, 2014.

[133] SLOBOGIN, C. «Principles of Risk Assessment: Sentencing and Policing», en *Ohio St. J. Crim. L.*, vol. 15, 2017. ; NILSSON, T. *et al.:* «The Precarious Practice of Forensic Psychiatric Risk Assessments», en *International Journal of Law and Psychiatry,* vol. 32, núm. 6, 2009, p. 406.

[134] BERK, R./ BLEICH, J.: «Forecasts of Violence…», *ob. cit.*, p. 87.

un contexto político social como el nuestro[135]. Es indudable, por tanto, que para una correcta valoración de si se pueden o no utilizar en las herramientas de valoración del riesgo datos como el género o la edad, no se pueden obviar otras consideraciones como son el alcance predictivo real de las herramientas y, a su vez, el tipo de afectación a los derechos que se deriva de las diferentes formas posibles de diagnóstico. Lo cual nos lleva al siguiente punto.

La segunda reflexión de interés se refiere a que si bien el alcance predictivo de estos algoritmos es limitado, especialmente en relación con decisiones jurídico-penales complejas, y puede estar sesgado, conviene recordar de dónde partimos: las herramientas predictivas de las que hablamos no vienen más que a hacer lo que ya se hacía y se hace a día de hoy de manera tradicional y manual y probablemente con los mismos sesgos o más, añadiendo, en algunos casos, una metodología más sistemática o científica[136]. Por expresarlo de otro modo: la decisión de por «dónde patrullamos», de «dónde ponemos los controles» e incluso la de «aplico o no la suspensión de la pena» se van a tener que seguir tomando, y no cabe duda alguna que todo aquello que aporte información científica y supere los prejuicios subjetivos mejorando la toma de decisiones debería ser visto como algo positivo. Pero esto debe tomarse con cautela. En primer lugar porque frente a la utopía de una IA futura imparcial que logre evitar que las diferentes agencias de seguridad y justicia puedan abusar sin control del poder inherente a su posición, lo que sabemos hasta el momento nos dice que los algoritmos, que reflejan con precisión nuestro mundo, parecen reflejar también nuestros prejuicios[137]. Sobre esto son numerosas las aportaciones que desde las ciencias de la computación ponen en duda la capacidad del aprendizaje automático propio de las arquitecturas de las IA actuales para el desarrollo de tareas clasificatorias o predictivas sin riesgo de discriminación de grupos vulnerables[138]. Más bien todo lo contrario, el sesgo humano puede perpetuarse e incrementarse mediante este tipo de técnicas, siendo esto especialmente grave en las IA aplicadas al sistema de justicia penal, ya que estos modelos van a depender, casi en su totalidad, del conjunto de datos de entrenamiento con los que elaboran sus modelos y estiman predicciones[139].

---

[135] Brennan-marquez, K./ henderson, S.: «Artificial Intelligence and Role-Reversible Judgment», en *Journal of Criminal Law and Criminology*, 2018, pp. 1-27.

[136] Para una revisión crítica, véase Brandariz Garcia, J. A.: *El modelo gerencial-actuarial de penalidad. Eficiencia, riesgo y sistema penal*, Dykinson, Madrid, 2016.

[137] Véase Isaac, W. S.: «Hope, Hype, …», *ob. cit.*, pp. 543 y ss.

[138] Para una revisión general de la literatura científica, véase Osoba, O. A, y Welser IV, W.: *An intelligence in our image: The risks of bias and errors in artificial intelligence*, Rand Corporation, 2017.

[139] Berk, R., Heidari, H., Jabbari, S., Kearns, M., y Roth, A.: «Fairness in Criminal Justice Risk Assessments:…», *ob. cit.*

En segundo lugar, a mi parecer es problemática la falsa expectativa sobre la absoluta (o cuanto menos, mejorada) fiabilidad de la predicción que da el concepto «Inteligencia» Artificial, que podría eliminar la realización por parte de los implicados de necesarios juicios de valor desde otras perspectivas de conocimiento y experiencia que pueden haber sido obviadas por la IA[140], todo ello teniendo en cuenta que algunos problemas son demasiado complejos para los datos de los que dispone la IA. Esto es lo que sucede, por ejemplo, con la cuestión de la valoración del riesgo de reincidencia, que exige análisis mucho más complejos que los que se pueden realizar con sistemas de IA como los que disponemos ahora a partir únicamente de datos demográficos, o con los algoritmos de detección de sujetos radicalizados a través de Internet, realizados casi en exclusiva por informáticos a partir de patrones matemáticos de palabras o metadatos separados de cualquier marco teórico sobre la conducta delictiva[141]. Frente a esto, la principal recomendación sería que todas estas IA deben ser creadas por equipos interdisciplinares que incluyan científicos sociales y juristas capaces de establecer tanto los criterios jurídicos y criminológicos de clasificación como de interpretación de resultados, y no exclusivamente por *data scientists* cuyo objetivo único sea la simplificación de la herramienta y la fiabilidad matemática de las inferencias[142].

Y queda una última reflexión. Las valoraciones que debemos hacer respecto a para qué puede ser utilizada en el sistema de justicia penal y para qué no una IA y respecto a qué tipo de variables pueden introducirse y cuáles no, deben llevarse a cabo teniendo en cuenta el muy diferente grado de afectación de derechos que puede derivarse de sus diferentes

---

[140] Y ello pese a que ya hay estudios empíricos recientes que han mostrado ventajas en el uso de IA frente a operadores humanos en materia de prevención de la delincuencia. Sírvanos como ejemplo el trabajo de Mohler, G. O., Short, M. B., Malinowski, S., Johnson, M., Tita, G. E., Bertozzi, A. L., y Brantingham, P. J.: «Randomized controlled field trials of predictive policing», en *Journal of the American Statistical Association*, vol. 110, n.º 512, 2015, pp. 1399-1411.

[141] Burnap, P., Williams, M. L.: «Cyber hate speech on twitter…», *ob. cit.*, pp. 223-242; Gao, L., Kuppersmith, A., y Huang, R.: «Recognizing Explicit and Implicit Hate Speech Using a Weakly Supervised Two-path Bootstrapping Approach», 2017. Disponible en: *arXiv preprint arXiv:1710.07394*.

[142] Para varios ejemplos de investigaciones interdisciplinares aplicados a la detección de discurso radical *online,* veáse Miro Llinares, F.: «La detección de discurso radical en Internet. Aproximación, encuadre y propuesta de mejora de los análisis de Big Data desde un enfoque de Smart Data criminológico», en Alonso Rimo, A., Cuerda Arnau, M.L., y Fernández Hernández, A. (dirs.), *Terrorismo, sistema penal y derechos fundamentales*, Tirant lo Blanch, Valencia, 2018, pp. 617-648; Miro Llinares, F., y Rodriguez-Sala, J. J.: «Cyber hate speech on twitter: Analyzing disruptive events from social media to build a violent communication and hate speech taxonomy», en *International Journal of Design & Nature and Ecodynamics*, vol. 11, n.º 3, 2016, pp. 406–415; Miró Llinares, F., Castro-Toledo, F. J., y Esteve, M.: «Dictionary of radicalisation. Developing a bag of word for its implementation into a classifying machine learning», en *American Society of Criminology Annual Meeting 2017,* American Society of Criminology, Filadelfia, 2017.

usos. Por expresarlo de otro modo, no es igual la incidencia en los distintos intereses dignos de tutela de una IA usada por la policía para intervenir las comunicaciones de una persona particular por la posible distribución de contenido radical terrorista donde se pueden ver afectados intereses como la tutela judicial efectiva, la presunción de inocencia, o la vida personal y familiar, sólo por citar algunos de los más evidentes; que de una IA que decide sobre la distribución de unidades policiales de patrullaje en diferentes áreas urbanas donde, a primera vista, la afectación es significativamente menor, pudiendo incidir, en casos muy específicos, por ejemplo, en la libertad deambulatoria de las personas. Y no es lo mismo, insisto, tanto por el alcance y consecuencias de la perfilación como por las garantías que habría que respetar en uno y otro caso. Aunque vamos hacia un escenario de utilización de herramientas de valoración del riesgo para, por ejemplo, decidir el tipo de sanción a aplicar o la aplicación o no de medidas cautelares como la prisión provisional, y pronto esas herramientas se automaticen por medio de IA, creo que antes de implantar tales sistemas resulta esencial discernir cual va a ser el alcance de tales instrumentos, que solo debiera ser informativo y nunca decisorio, y cómo debiera configurarse las herramientas con respecto a las variables que incluya en cada uno de los casos. En esta línea, de hecho, ha ido el propio Parlamento Europeo en 2017, al manifestar respecto a los modelos de actuación policial predictiva que no es lo mismo la realización de predicciones probabilísticas sobre lugares o acontecimientos, que la perfilación individual o predicción realizada sobre personas particulares[143]. Es obvio que las garantías para evitar los sesgos en estos casos deben ser mayores dado que los riesgos de discriminación por el uso de este tipo de IA es mayor.

No será este el trabajo en el que dé por cerradas todas las cuestiones aquí planteadas y menos las muchas que quedan por plantear. Pero habrá que empezar a dar pasos para hacerlo, pues es obvio que las IA han logrado ser incorporadas con cierto éxito a tareas de toma de decisiones policiales o judiciales, y no hay vistas de que su actividad merme en un futuro, sino más bien todo lo contrario. Frente a ello podemos adoptar: una visión conformista, y reconocer que el uso de tales IA nos traerá más ventajas que inconvenientes y dejar estos de lado para abrazar sus beneficios; una visión negativa, que renuncie a las evidentes ventajas que el uso de esta IA puede traer ante las también claras discriminaciones que producirá; y una visión realista, que acepte lo que la IA puede traernos de incorporación de evidencias científicas para la mejora de la prevención del crimen a la par que reconozca las enormes limitaciones que tales herramientas aún tienen y que les puede impedir

[143] Más específicamente, el punto 29 de la resolución (2016/2225 (INI)) *on fundamental rights implications of Big data: privacy, data protection, non-discrimination, security and law enforcement*.

aún ser operativas bajo premisas equitativas y comprenda los enormes riesgos que un mal uso puede conllevar. Si adoptamos esta visión reconoceremos que más que potenciar una IA sin control debemos potenciar el binomio persona-máquina para mejorar la información disponible para la primera gracias a la capacidad de cálculo de la segunda, pero siempre desde la consideración de los aspectos metodológicos y técnicos que se esconden detrás de los algoritmos discriminatorios que deberían ser minimizados en arquitecturas más éticas basadas en estrategias metodológicas y diseños que reduzcan los riesgos aquí planteados.

En este sentido, y en la línea de lo que ya han señalado otros autores, hay que admitir que, excepto en casos triviales, va a ser imposible maximizar, al mismo tiempo, la exactitud de la herramienta y su absoluta imparcialidad[144]. La realidad no es imparcial, por lo que para poder evaluar adecuadamente tendremos que aceptarlo, y si lo que queremos es que el juez disponga de toda la información necesaria para realizar las evaluaciones más justas y adecuadas tendremos que asumir la utilización de toda la información posible, incluso de aquella que no está directamente relacionada con el sujeto. Pero si lo que vamos a hacer con estas herramientas, por ejemplo, es determinar una posible privación de su libertad, creo que la información global que se le debe ofrecer por medio de la IA al órgano judicial debe permitir al juez centrarse en los aspectos individuales del sujeto pese a que incluyan toda la información disponible, por ejemplo permitiéndole conocer como cambia la valoración del riesgo por la IA en el caso de que se elimine la condición problemática. Porque, si en un futuro próximo se opta por la utilización de herramientas actuariales de IA para la valoración del riesgo deberíamos estar seguros, como ha señalado SLOBOGIN[145], de que las mismas pudieran cumplir tres exigencias mínimas: adecuación a lo que efectivamente necesita saber el juez que va a tomar la decisión; validez conforme a los parámetros científicos; y equidad, lo cual, a mi parecer, exige al menos que el juez pueda, de toda la información disponible, eliminar aquellas variables que no resulten justas conforme a lo que está siendo evaluado, y valorar el riesgo de la persona por lo que ella ha hecho o puede hacer.

La incorporación de cualquier tecnología debe estar basada en algo volitivo, y no sólo en lo posibilístico. Debemos incorporar la IA a la justicia penal y a la actividad policial no porque podemos hacerlo sino porque gracias a ello podamos mejorar la práctica de la justicia penal[146].

---

[144] BERK, R., HEIDARI, H., JABBARI, S., KEARNS, M., y ROTH, A.: «Fairness in Criminal Justice Risk Assessments:…», *ob. cit.* pp. 27.

[145] SLOBOGIN, C. «Principles of Risk Assessment: Sentencing and Policing», en *Ohio St. J. Crim. L.*, vol. 15, 2017, pp. 594 y ss

[146] En el mismo sentido, en varios momentos del libro, NIEVA FENOLL, J.: Inteligencia artificial y proceso …, *ob. cit.*

Está claro que no se puede esperar de ninguna IA que sea capaz de predecir el futuro. Es evidente, como han dicho Berk y otros, que ninguna herramienta de valoración del riesgo será capaz de revertir siglos de injusticia racial o de desigualdad de género[147]. Pero creo que si comprendemos su auténtico alcance, incluimos la visión ética y jurídica en un quehacer en el que predominan los científicos sociales y los de la computación y, además, adoptamos las cautelas necesarias fruto del respeto a nuestros derechos y principios, deberíamos ser capaces, gracia a la IA, de hacer justicia un poco mejor.

[147] BERK, R., HEIDARI, H., JABBARI, S., KEARNS, M., y ROTH, A.: «Fairness in Criminal Justice Risk Assessments:…», *ob. cit.* pp. 32 y ss.